2017

Nonparametric regression models with and without measurement error in the covariates, for univariate and vector responses: a Bayesian approach

Eduardo Antonio Trujillo-Rivera
Iowa State University

Nonparametric regression models with and without measurement error in the covariates for univariate and vector responses: A Bayesian approach

by

Eduardo Antonio Trujillo Rivera

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Alicia L. Carriquiry, Co-major Professor
Daniel J. Nordman, Co-major Professor
Kris De Brabanter, Co-major Professor
Jarad B. Niemi
Stephen B. Vardeman

The student author and the program of study committee are solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2017

DEDICATION

I would like to dedicate this dissertation to my parents María del Carmen Rivera-Cruz and Antonio Trujillo-Narcía, because of the love they have shared with me, for all their unconditional support. I dedicate to them this work because without them, I would have not arrived to the place I am now. They provided me with basic principles of life, with my first years of formation and of education; they taught me with their example, perseverance, passion, and the love for learning.

I dedicate this dissertation to my brother Alejandro Trujillo-Rivera because we have shared so many experiences that forged my form of being and my for of behave. He has listened to me every time I needed it, for being my friend.

To them, I dedicate this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research, the writing of this dissertation and the support needed during these five and half years in Ames, Iowa.

First and foremost, to my three advisors Dr. Alicia L. Carriquiry, Dr. Daniel J. Nordman and Dr. Kris De Brabanter. I thank you for all the weekly meetings we hold during the last four years, for your guidance, your patience, your invaluable support throughout this research, and for the suggestions and corrections during the process of writing of the dissertation. Your insights and words of encouragement have often inspired me and have pushed me to achieve the completion of my graduate education.

I want to thank Dr. Jarad B. Niemi and Dr Stephen B. Vardeman for their invaluable observations to my research.

I would also like to thank to all my professors, who shared with me their knowledge, their philosophy of life and experience. I learned so much from them.

I thank to all my friends that shared with me all those mornings and evenings in the classrooms, all those days and nights studying, sharing with me their knowledge, practicing, solving problems, discussing, learning and trying to comprehend so many topics. Sometimes, I wanted to finish the day even when there was still so much to do; their presence and enthusiasm pushed me to continue a lot more.

I want to thanks to the staff in the Department of Statistics that helped me in numerous occasions.

I thank to the Department of Statistics and to the Iowa State University because all the intellectual and economic support I received from them.

ABSTRACT

This dissertation addresses the problem of estimation in multivariate non-parametric regression of real value and vector valued functions when there is classical measurement errors in the covariates. Different estimation approaches, including selection of bandwidth parameters, are studied first and compared for the case of no measurement error, and then for the error case. New theoretical results related to criteria for selecting the bandwidth parameter are presented for the vector valued regression problem. We also conjecture on possible extensions of the methods to improve estimation in the multivariate response case.

In the context of semi-parametric regression with multiple covariates, it is known that the solution to the penalized least squares minimization problem can be interpreted as the mean of the posterior distribution arising in the context of an empirical Bayesian approach. The probability model in this approach has a Gaussian process as prior on the target regression function with co-variance structure depending on the reproducing kernel of an associated reproducing kernel Hilbert space. By the Representer Theorem, the solution to the minimization problem can be expressed as a linear combination of a set of known basis functions. We prove that under different a Bayesian model with multivariate normal priors on the coefficients and covariance structure depending on a reproducing kernel, it is possible to obtain the same posterior estimates of the regression function as with the previous formulation with the Gaussian process prior. Our approach has an advantage over its predecessor; to predict the value of the target function on any domain and to produce credible intervals for the predictions, we only need to evaluate known basis functions using estimated parameters. In contrast, when using the previous Bayes formulation with Gaussian process prior, we first need to fix the points where the Gaussian process is to be estimated but subsequent evaluations of the process is done externally; for computational reasons, obtaining an exact solution to the penalized least square minimization problem is not practical; instead, we review, modify and implement an

approximate solution. We show that the full conditional posterior distribution of the point-wise regression estimates is the same in both approaches.

We evaluated the performance of our method using simulation. We compared our Bayesian approach applied to existing methods proposed for estimation in non-parametric regression in the frequentist setting, including thin plate splines, a linear mixed model interpretation of thin plate splines, and tensor product splines with marginal thin plate splines. In all cases, we computed the previously mentioned approximate solution to the optimization least square problem. The computation of smoothing parameters is done via empirical Bayes approach that involves the minimization of score functions. We considered three different score functions from the literature. The linear mixed model formulation enables us to write the smoothing parameter as the ratio of two variances and therefore we can estimate the parameter, as a fourth approach, using the standard Bayesian estimation framework. We compare the various approaches by focusing on frequentist properties of the Bayesian estimator of the regression function and of the point-wise credible intervals. In particular, we compute average coverage rates of the credible intervals for all methods, where the average is taken over the prediction points. We find that the average coverage probability is close to the nominal level, at least for predictions inside the observation region for the covariates; while point-wise credible intervals are not to be trusted to have nominal coverage, unless they are inside the region of covariate observation and only when using specific methods to select smoothing parameters.

The simulation has two objectives: to study the performance of the estimators and to examine potential approaches involving basis functions with tractable form which might be used in a more complex setting with errors in the measurements. We argue that the Bayesian framework applied to the thin plate spline approach is an acceptable trade off between computational complexity required to fit and predict from the model and the frequentist properties of the estimators. Using the proposed Bayesian model and the thin plate splines, we extend our Bayes model for the regression problem with multiple regressors and classical measurement error in the covariates. We carried out similar simulation study with the purpose of studying the frequentist properties of the estimators. We discuss simulation results that refer to point-wise estimation of the regression function, empirical coverage of point-wise credible intervals for

evaluations of the regression function, and to performance of estimators of the observation-error variance.

While reviewing the literature, we found that many results are either presented without proof or with proofs that seemed incomplete to us. In those cases, we endeavored to write complete proofs for those results on which we relied. If the proof of a proposition is presented in this dissertation, that indicates that it was not available in the literature and can be considered original research. Whenever a proposition is listed without a proof, it means that the proof was published elsewhere and we include the corresponding citation.

Finally, we also consider the case where the response is vector-valued and the form of the mean regression function is unknown. We first propose an approach of estimation when there is no measurement error in the covariates. We then extend the method to the case where covariates are measured with classical error. As in the univariate response case, we do not assume a form of the regression function but we do formulate a set of assumptions that must be met. We propose – without complete proof – three methods for computing the smoothing parameters and extend the methods to theoretically address calculation of a diagonal bandwidth matrix and a general bandwidth matrix. We illustrate these methods via simulated examples.

CHAPTER 1. OVERVIEW

Let $\mathbb{X}$ be a non-empty set and $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{X} \times \mathbb{R}$ denote a sample of regressors $\mathbf{x}_i$ and response variables y_i. Let $\mathcal{H} \subset \{\eta : \mathbb{X} \to \mathbb{R}\}$ be a class of response curves for describing the response mean as a function of the regressors; in particular $\mathcal{H}$ will be a Reproducing Kernel Hilbert Space, *RKHS* (Definition 35) possibly of infinite dimension, dimension in the sense of vector linear space over $\mathbb{R}$. Let be J the square of the norm in $\mathcal{H}$. The penalized least squares minimization problem can be formulated as follows: find the function $\eta \in \mathcal{H}$ that solves

$$\arg\min_{\eta \in \mathcal{H}} \sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i))^2 + n\lambda J(\eta) \tag{1.1}$$

with penalty parameter $\lambda > 0$. This problem has been widely studied by different authors using theory of linear operators in Hilbert spaces; see (Weidmann (1980); Akhiezer and Glazman (1981a,b)) from the perspective of *RKHS's* , and (Aronszajn (1950); Duchon (1977); Wahba and Craven (1978); Kimeldorf and Wahba (1971); Meinguet (1979); Wahba and Wendelberger (1980); Wahba (1985, 1987); Chen (1993); Wood (2003); Gu (2013)).

The choice of appropriate smoothing parameters λ and other hidden smoothing terms inside the penalty term J is challenging and has been approached by defining score functions that need to be minimized. Examples of approaches that minimize a score function include unbiased parameter estimation that minimizes some loss function (Mallows (1973) (UERL)), cross validation and generalized cross validation (Wahba and Craven (1978) (GCV)), and scores obtained in the context of maximum likelihood estimation in certain models (Wecker and Ansley (1983), Wahba (1985),Li (1986) (RML)). When there are no smoothing parameters in J, λ may be interpreted as the ratio of two variance components in a linear mixed effects model (Ruppert et al. (2003)) equivalent to (1.1); such linear mixed model is obtained from a perspective of the best linear unbiased prediction (BLUP) (Henderson (1973), Robinson (1991)).

The solution to the problem (1.1) is by definition in $\mathcal{H}$ which can in principle be a space of functions of infinite dimension. Under certain conditions on J, it can be shown that a unique solution exists in a subspace of $\mathcal{H}$ that is completely and uniquely described by a finite number of functions; a finite dimensional space. This result is known as the *Representer Theorem* (Schölkopf et al. (2001)) and plays a key role in the rest of thesis. The advantage of knowing that the unique solution belongs to a fully described finite dimensional space is that the infinite dimensional problem (1.1) is reduced to a finite dimensional problem where only a finite number of parameters that need to be estimated. For given smoothing parameters, the general approach for solving (1.1) is to write the solution to (1.1) as a finite linear combination of a known basis set of functions that depend entirely on the sampling points $\{\mathbf{x}_i\}_{i=1}^n$ and on the associated reproducing kernel of $\mathcal{H}$. Using this finite representation, one can transform the functional minimization problem (1.1) into a simpler one: to solve a real linear finite system of equations. This system may still be very large and computationally intensive, but at least might be approximately solved and the rate of approximation can be studied.

For computational reasons, we need to find an approximate solution to (1.1) (Gu and Kim (2002), Wood (2003), Kim and Gu (2004)) and one way to do it is by solving a lower dimensional linear system. This solution has the same asymptotic convergence rate as the exact solution. Some loss in the accuracy of point estimates of the function η is the price we pay for a fast algorithm that can be used in practice.

The penalty term J reflects, up to the value of $n\lambda$, the smoothness desired for the solution to (1.1). In practice we must choose J in a case by case basis. In general, for a smooth solution (in the sense of continuity of derivatives) we can choose J to be in the family of the thin plate splines (Duchon (1977), Wahba and Wendelberger (1980), Wood (2003), Ruppert et al. (2003), Gu (2013)) or we can construct new penalty terms using a tensor product of Hilbert spaces which leads to solutions to (1.1) as tensor product smoothing splines (Barry et al. (1986),Wahba (1987), Gu and Wahba (1993a), Gu and Wahba (1993b), Chen (1993) , Barry et al. (1986), Gu (2013)).

In addition to point estimation, we wish to obtain confidence intervals for the point estimates of the regression function. This can be accomplished by proposing Bayesian estimation methods

for regression parameters with the property that the mean of the full conditional distribution of the estimated function is either the approximate solution, or the exact solution to the penalized least square minimization problem (1.1). The first attempt to fit smoothing splines within a Bayesian framework is by (Kimeldorf and Wahba (1970)); the selection of the smoothing parameters is accomplished with the help of the sampling points $\{\mathbf{x}_i\}_{i=1}^n$, an empirical Bayes approach. Further work was developed by (Wahba (1978)) for the simple regression case. Other relevant references are (Wahba (1983)), (Berry et al. (2002)) and (Kim and Gu (2004)).

Here we review the theory that we need, to obtain the exact and the approximate solutions to the penalized least square minimization problem in (1.1). We also review four existing methods to calculate smoothing parameters, and the theory underlying empirical Bayesian methods that we use to obtain credible bands around our estimates. We found that many results are presented in the literature without proof. Whenever we could not find a proof for a proposition on which we relied, we endeavored to produce the proof. Therefore, an additional contribution of this work is the proof of all the results for which we could not find one in the literature. All the theoretical background results are given in appendix *B*. The general rule is the following: if the proposition has been proven, we provide the corresponding reference, and if it has not (or if the proof is not easily found in the literature) we include a detailed proof in this appendix.

The rest of the dissertation is organized as follows. In Chapter 2 we review the concept of reproducing kernel Hilbert Spaces (*RKHS*), the space in which we develop our method. We also revisit existing results on the form, uniqueness, and interpretation of the exact solution to the penalized least square estimation problem. In Section 2.1 we provide details about the properties of the exact solution to (1.1) for the case of the thin plate splines, tensor splines with thin plate splines as marginals, and linear mixed model interpretation of the exact solution to the thin plate splines. We give an explicit expression of the reproducing kernel in the thin plate splines setting and discuss a computational algorithm to evaluate it in Sections 2.1.1, 2.1.1.1 and 2.1.1.2. The theory of tensor thin plate splines is reviewed in Section 2.1.2, where we also describe computational approaches to their application. The linear mixed model framework is described in Section 2.1.3. The results obtained when we focus on computing the approximated

solution to the penalized minimization problem are shown in Section 2.2, first for smoothing splines (Section 2.2.1) and then for the linear mixed model representation (Section 2.2.2). To approximate the solution to (1.1) we must select a set of knots; a criterion for selecting the knots is described in Section 2.2.3 and in Section 2.2.4 we discuss a method to select the degree of the smoothness provided by the penalty J of the thin plate splines. Algorithms to compute the smoothing parameter(s) are described in Section 2.3.

In Chapter 3 we propose a novel, fully Bayesian approach to solve the penalized minimization problem on which we focus. Initially, we assume that the bandwidth parameter(s) is given. The proposed approach has the property that the full posterior distribution of the mean regression function is the approximated solution to (1.1) and equals the regression estimator from Kim and Gu (2004). Both methods are similar in terms of computational effort. The method we propose, in contrast to Kim *et.al.*'s, can be extended to the measurement error case in a straightforward way (Section 4). Kim *et.al.* treat the regression function as a Gaussian process while we treat it as a linear combination of a set of known fixed basis functions. When using Kim*et.al.*'s method in the measurement error case, prediction of η for new values of the covariates $\{\chi_i\}_{i=1}^N$ is computationally challenging because the unobserved values of the covariates $\{\mathbf{x}_i\}_{i=1}^n$ need to be estimated; the estimation is carried out using an MCMC algorithm. The estimates $\{\hat{\mathbf{x}}_i\}_{i=1}^n$ change in each iteration, which means that we need to keep track of the sequence of estimates of $\{\eta(\chi_i)\}_{i=1}^N$ on the grid $\{\chi_i\}_{i=1}^N$ selected in each step. This then means that Kim *et.al.* must choose from the beginning all the points at which they wish to predict the value of the regression function, and they also need to save all those estimates. With the method we propose we only need to save the estimates of the coefficients in the regression model so that the fixed basis functions can be evaluated at $\{\chi_i\}_{i=1}^N$ to produce an estimate of $\{\eta(\chi_i)\}_{i=1}^N$. The estimates are obtained from the posterior predictive distribution. In Section 4 we apply our method in the case where covariates are subject to classical measurement error. The computational effort is big, which means that it is important to streamline calculations as much as we can without sacrificing accuracy. We provide empirical justification for some computational short-cuts by showing the results of a simulation study in Section 3.3.

We address the problem of non-parametric estimation in a multivariate regression model with measurement error in the covariates in Section 4. To do so, we extend the Bayesian approach from Section 3. We again evaluate the performance of the estimator in terms of estimation bias and coverage of credible intervals by simulation, using a design for the simulation study that closely parallels the one we have already described. Here, we also compare results we obtain when accounting for the measurement error and when using the average of a large number of replicates of the covariates associated with an experimental unit as if it was the true (unobservable) value.

Finally, Chapter 5 refers to the estimation problem when the response is vector-valued and the elements of the response vector are correlated. We first focus on the case where covariates are observed with no error using tools similar to those described in Section 3 for univariate responses. As in Section 3, we propose a Bayesian model that results in a posterior distribution for $\eta(\mathbf{x}_i)$ whose mean can be interpreted as the solution to

$$\arg\min_{\eta\in\mathcal{H}} \sum_{i=1}^{n} \|\eta(\mathbf{x}_i) - \mathbf{y}_i\|_{\mathbb{Y}}^2 + \lambda_A \|\eta\|_{\mathcal{H}}^2.$$

The regression with measurement error in the covariates is resolved by extending the Bayesian model developed for the error-free case as we did in Section 4. Again we propose three methods to choose bandwidth parameters and we make conjectures about a methodological extension that may enable choice of diagonal and general bandwidth matrices. We illustrate these methods using a few examples.

Appendix *A* contains definitions. In Appendix *B*, we provide proofs for results from the literature for which we could not find one. In Appendix *C*, we show figures with more detailed results than the ones shown in the main body of this document. Finally, in Appendix *D* we discuss computational challenges and offer a list of good practices for anyone who wishes to implement this type of estimation method.

CHAPTER 2. PRELIMINARIES: SMOOTHING SPLINES MULTIPLE REGRESSION WITH NO ERROR IN THE COVARIATES

Let $\mathbb{X}$ be a non-empty set and $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{X} \times \mathbb{R}$ denote a training set. Let $\mathcal{H}$ be a reproducing kernel Hilbert space denoted *RKHS* which can be written as tensor sum decomposition (Definition 40) $\mathcal{H} = \bigoplus_{i=0}^p \mathcal{H}_i \subset \{\eta : \mathbb{X} \to \mathbb{R}\}$ with $\mathcal{H}_i \subset \mathcal{H}$ closed sub-spaces which may be independent, and have inner products $(f_i, g_i)_i$ and reproducing kernels R_i. By Theorem 47, there exists a unique $f_i \in \mathcal{H}_i$, the projection of f onto $\mathcal{H}_i$. For convenience we write $(f_i, g_i)_i = (f, g)_i$. A general form of a penalized least squares functional in $\mathcal{H}$ is written as

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \eta(\mathbf{x}_i))^2 + \lambda J(\eta), \tag{2.1}$$

with penalty term

$$J(\eta) = J(\eta, \eta) = \sum_{i=1}^{p} \theta_i^{-1} (\eta, \eta)_i , \tag{2.2}$$

where it is assumed that $\eta \in \mathcal{H}$ and $\mathcal{H}$ is a space whose elements have the properties that are needed for the computation of J with $J(\eta) < \infty$. As shown later, J is the squared norm induced by the inner product in $\mathcal{H}$.

The parameter λ controls the trade-off between smoothness as measured by J, and the discrepancy between the fitted function and the training set as measured by the quadratic loss function $\sum_{i=1}^n (y_i - \eta(\mathbf{x}_i))^2$. The positive tunning parameters $\{\theta_i\}_{i=1}^p$ allow for re-scaling of the metrics $(\cdot, \cdot)_i$. The smoothing parameters λ and $\{\theta_i\}_{i=1}^p$ need to be selected (see Section 2.3). The penalty term J is overparametrized in the sense that we only need to specify the ratios λ/θ_i. We may choose to fix λ, or to fix one θ_i, or to impose a constraint on λ and on $\{\theta_i\}_{i=1}^p$. For easier interpretation, it is desirable to maintain the symmetry of the penalty term (2.2). For this dissertation, when $p = 1$ in equation (2.1) and we set $\theta_1 = 1$, this corresponds to the

case of the thin plate splines (Section 2.1.1), while $p > 1$ corresponds to the case of tensor thin plate splines (Section 2.1.2). To avoid identifiability issues, we can set $\lambda = 1$.

The bi-linear form $J(f,g) = \sum_{i=1}^{p} \theta_i^{-1} (f,g)_i$ is assumed to be an inner product in $\bigoplus_{i=1}^{p} \mathcal{H}_i = \mathcal{H} \ominus \mathcal{H}_0$ which has a reproducing kernel

$$R_J = \sum_{i=1}^{p} \theta_i R_i.$$

R_J is the reproducing kernel because it has the reproducing property, for $\mathbf{x} \in \mathbb{X}$:

$$\begin{aligned} J(R_J(\mathbf{x},\cdot), f) &= \sum_{i=1}^{p} \theta^{-1} \left(\left[\sum_{j=1}^{p} \theta_j R_j(\mathbf{x},\cdot) \right]_i, f_i \right)_i = \sum_{i=1}^{p} \theta^{-1} (\theta_i R_i(\mathbf{x},\cdot), f_i)_i \\ &= \sum_{i=1}^{p} (R_i(\mathbf{x},\cdot), f_i)_i = \sum_{i=1}^{p} f_i(\mathbf{x}) = f(\mathbf{x}). \end{aligned}$$

The reproducing kernel is unique by Theorem 56. The null space $\mathcal{N}_J = \mathcal{H}_0$ of J should have finite dimension l, which is important for the existence of the minimizer as described in Theorem 50 and in its proof; the proof invokes a recurrent argument where convergence is ensured only if the dimension of $\mathcal{N}_J$ is finite. If the conditions of Theorem 50 are met, then for fixed λ and $\{\theta_i\}_{i=1}^{p}$, the Representer Theorem (Theorem 52), (Kimeldorf and Wahba (1971); Wahba and Wendelberger (1980); Schölkopf et al. (2001)), states that the solution $\eta_\lambda \equiv \eta$ to (1.1) can be expressed as

$$\begin{aligned} \eta(\mathbf{x}) &= \sum_{i=1}^{l} d_i \psi_i(\mathbf{x}) + \sum_{i=1}^{n} c_i R_J(\mathbf{x}_i, \mathbf{x}) \\ &= \psi(\mathbf{x})^\intercal \mathbf{d} + \xi(\mathbf{x})^\intercal \mathbf{c}, \end{aligned} \tag{2.3}$$

where $\{\psi_\nu\}_{\nu=1}^{l}$ is a basis of the space $\mathcal{N}_J = \mathcal{H}_0$ and

$$\begin{aligned} \mathbf{d} &= (d_1 \cdots d_l)^\intercal, d_i \in \mathbb{R}, \\ \mathbf{c} &= (c_1 \cdots c_n)^\intercal, c_i \in \mathbb{R}, \\ \psi(\mathbf{x}) &= (\psi_1(\mathbf{x}) \cdots \psi_l(\mathbf{x}))^\intercal, \\ \xi(\mathbf{x}) &= (R_J(\mathbf{x}_1, \mathbf{x}) \cdots R_J(\mathbf{x}_n, \mathbf{x}))^\intercal. \end{aligned}$$

Plugging (2.3) in (2.1) and using properties of the reproducing kernel R_J (Proposition 53), the problem of minimizing (2.1) in the space $\{\eta \in \mathcal{H} : J(\eta) < \infty\}$ is reduced to minimize

$$(\mathbf{y} - S\mathbf{d} - Q\mathbf{c})^\intercal (\mathbf{y} - S\mathbf{d} - Q\mathbf{c}) + n\lambda \mathbf{c}^\intercal Q \mathbf{c} \tag{2.4}$$

with respect to $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{d} \in \mathbb{R}^l$, where $Q \in \mathcal{M}_{n\times n}(\mathbb{R})$ with $(i,j)th$ entry $R_J(x_i, x_j)$ and $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ with $(i,j)th$ entry $\psi_j(x_i)$. By differentiating (2.4) with respect to $\mathbf{c}$ and $\mathbf{d}$ and then equating the derivatives to $\mathbf{0}$ with the help of Lemma (64), we obtain a linear system 2.5 whose solution provides the values $\mathbf{d}$ and $\mathbf{c}$ that are inflection points

$$\begin{aligned} Q\{(Q+n\lambda I)\mathbf{c} + S\mathbf{d} - \mathbf{y}\} &= \mathbf{0} \\ S^{\intercal}\{Q\mathbf{c} + S\mathbf{d} - \mathbf{y}\} &= \mathbf{0}. \end{aligned} \tag{2.5}$$

We now establish that there is a unique solution to (1.1). By Proposition 48, the functional $\mathsf{L}(\eta) := \sum_{i=1}^{n}(y_i - \eta(\mathbf{x}_i))^2$ is continuous and convex in $\mathcal{H}$ and when S is of full column rank, the convexity is strict in $\mathcal{N}_J$ and L has a minimizer in this space. Theorem 50 states that a minimizer of (2.1) exists as long as L has a minimizer in $\mathcal{N}_J$. Proposition 48 states as well that (2.1) is strictly convex in $\mathcal{H}$ when S is of full column rank and by Proposition 49, we find that (1.1) has a unique solution on $\mathcal{H}$.

Theorem 51 enables the interpretation of the minimization problem (1.1) as an estimation problem. It states that, for a given smoothing parameter $\lambda > 0$, minimizing (2.1) in $\mathcal{H}$ is equivalent to finding the function $\hat{\eta} \in H$ that best fits the training data in the sense of minimizing the quadratic loss function L, where $\hat{\eta}$ is subject to the constraint that $\hat{\eta}$ is in the ball of radius ρ ($J(\hat{\eta}) \leq \rho^2$); ρ depending on both λ and the Gâteaux derivative of L.

Even when the unique solution to (1.1) exists, if Q is not of full column rank, then (2.4) may have multiple minimizers with $\mathbf{c}$ and $\mathbf{d}$ satisfying (2.5), all of them yield the same estimate of η_λ by (2.3). In order to obtain a unique solution in practice, it is usually assumed that $(Q+n\lambda I)\mathbf{c} + S\mathbf{d} - \mathbf{y} = 0$. The new system

$$\begin{aligned} Q\{(Q+n\lambda I)\mathbf{c} + S\mathbf{d} - \mathbf{y}\} &= 0 \\ S^{\intercal}\{Q\mathbf{c} + S\mathbf{d} - \mathbf{y}\} &= 0 \\ (Q+n\lambda I)\mathbf{c} + S\mathbf{d} - \mathbf{y} &= 0 \end{aligned} \tag{2.6}$$

is consistent regardless of the rank of Q. The system (2.6) is equivalent to

$$\begin{aligned} Q\{0\} &= 0 \\ S^{\intercal}\{-n\lambda\mathbf{c}\} &= 0 \end{aligned}$$

$$(Q + n\lambda I)\,\mathbf{c} + S\mathbf{d} - \mathbf{y} = 0,$$

which is equivalent to

$$\begin{aligned} S^\intercal \mathbf{c} &= 0 \\ (Q + n\lambda I)\,\mathbf{c} + S\mathbf{d} &= \mathbf{y}. \end{aligned} \tag{2.7}$$

The system of equations (2.7) has a unique solution and it corresponds to the solution to (2.5). Assume that S has full column rank, an assumption required for the existence and uniqueness of the minimizer of (2.1). The QR-decomposition (Golub and Van Loan (2012)) of S can be written as

$$S = FR^* = (F_1, F_2)\begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} = F_1\tilde{R}, \tag{2.8}$$

where F has orthogonal columns as measured by the euclidean inner product, $\tilde{R}$ is a upper triangular matrix, and F_1 and $\tilde{R}$ are unique because S is of full column rank. Using (2.8), the first equation in (2.7) and the fact that $\tilde{R}$ is square non-singular we get

$$F_1^\intercal \mathbf{c} = 0. \tag{2.9}$$

On the other side of (2.8), since F_2 has orthogonal columns, it holds that

$$I_n = F_2 F_2^\intercal. \tag{2.10}$$

Multiplying from the left the second equation in (2.7) by $F_2^\intercal$ we can solve for $\mathbf{c}$ as follows:

$$F_2^\intercal (Q + n\lambda I)\,\mathbf{c} + F_2^\intercal S\mathbf{d} = F_2^\intercal \mathbf{y}$$

$$F_2^\intercal (Q + n\lambda I)\,\mathbf{c} = F_2^\intercal \mathbf{y} \tag{2.11}$$

$$F_2^\intercal (Q + n\lambda I)\, F_2 F_2^\intercal \mathbf{c} = F_2^\intercal \mathbf{y} \tag{2.12}$$

$$(F_2^\intercal Q F_2 + n\lambda F_2^\intercal F_2)\, F_2^\intercal \mathbf{c} = F_2^\intercal \mathbf{y} \tag{2.13}$$

$$F_2^\intercal \mathbf{c} = (F_2^\intercal Q F_2 + n\lambda F_2^\intercal F_2)^{-1} F_2^\intercal \mathbf{y} \tag{2.14}$$

$$\mathbf{c} = F_2 (F_2^\intercal Q F_2 + n\lambda I)^{-1} F_2^\intercal \mathbf{y}, \tag{2.15}$$

where (2.11) follows from $F_2^\intercal S = F_2^\intercal F_1 \tilde{R} = 0\tilde{R} = 0$, (2.12) follows from (2.10), (2.13) is obtained by simple algebra, (2.14) is obtained by pre-multiplying by the inverse of $F_2^\intercal Q F_2 + n\lambda I$ and

(2.15) results by pre-multiplying by F_2 and using (2.10). Similarly we can obtain $\mathbf{d}$ by pre-multiplying the second equation in (2.7) by $F_1^\intercal$ in the following way:

$$\begin{aligned} F_1^\intercal Q\mathbf{c} + n\lambda F_1^\intercal \mathbf{c} + F_1^\intercal S\mathbf{d} &= F_1^\intercal \mathbf{y} \\ F_1^\intercal Q\mathbf{c} + F_1^\intercal S\mathbf{d} &= F_1^\intercal \mathbf{y} && (2.16) \\ F_1^\intercal S\mathbf{d} &= F_1^\intercal \mathbf{y} - F_1^\intercal Q\mathbf{c} && (2.17) \\ \tilde{R}\mathbf{d} &= F_1^\intercal \mathbf{y} - F_1^\intercal Q\mathbf{c} && (2.18) \\ \mathbf{d} &= \tilde{R}^{-1} F_1^\intercal \left(\mathbf{y} - Q\mathbf{c}\right) && (2.19) \end{aligned}$$

where equation (2.16) follows from (2.9), (2.17) and (2.19) are the result of simple algebra and (2.18) is obtained using equation (2.8) and $F_1^\intercal F_1 = I_n$. Therefore, we find that

$$\begin{aligned} \hat{\mathbf{c}} &= F_2 \left(F_2^\intercal Q F_2 + n\lambda I\right)^{-1} F_2^\intercal \mathbf{y} \\ \hat{\mathbf{d}} &= \tilde{R}^{-1} F_1^\intercal \left(\mathbf{y} - Q\mathbf{c}\right). \end{aligned} \tag{2.20}$$

The minimizer $\hat{\eta}$ to (1.1) is then $\hat{\eta}(\mathbf{x}) = \psi(\mathbf{x})^\intercal \hat{\mathbf{d}} + \xi(\mathbf{x})^\intercal \hat{\mathbf{c}}$, (2.3). Denoting the fitted values by $\hat{\mathbf{y}} := \hat{\eta}(\mathbf{x})$ some algebra leads to

$$\begin{aligned} \hat{\mathbf{y}} &= S\hat{\mathbf{d}} + Q\hat{\mathbf{c}} \\ &= \left(F_1 F_1^\intercal + F_2 F_2^\intercal Q F_2 \left(F_2^\intercal Q F_2 + n\lambda I\right)^{-1} F_2^\intercal\right) \mathbf{y} \\ &= \left(I - F_2 \left(I - F_2^\intercal Q F_2 \left(F_2^\intercal Q F_2 + n\lambda I\right)^{-1}\right) F_2^\intercal\right) \mathbf{y} \\ &= \left(I - n\lambda F_2 \left(F_2^\intercal Q F_2 + n\lambda I\right)^{-1} F_2^\intercal\right) \mathbf{y}. \end{aligned}$$

The symmetric matrix

$$A(\lambda) = I - n\lambda F_2 \left(F_2^\intercal Q F_2 + n\lambda I\right)^{-1} F_2^\intercal \tag{2.21}$$

is the smoothing matrix associated with the minimization problem (1.1). The expression for $A(\lambda)$ depends on $\{\theta_i\}_{i=1}^p$ defined in (2.2) and hidden in Q, but to simplify notation, we omit this dependence. The smoothing matrix can be alternatively written (Corollary 55) as

$$A(\lambda) = I - n\lambda M^{-1} \left(I - S \left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1}\right), \tag{2.22}$$

where $M = Q + n\lambda I$. Expressions (2.21) and (2.22) will come in handy in the next sections.

2.1 Non-Parametric Regression Models

In this section we discuss various non-parametric estimation approaches for non-parametric regression models. We provide some theoretical background and implementation practices in the following sub-sections.

2.1.1 Thin plate splines

Thin plate splines are the generalization of cubic splines to any dimensions and can be used to obtain a smooth estimate of a surface by data interpolation and smoothing. In particular, they can be applied within the framework described in Section 2.

Define a semi-inner product (Wahba and Wendelberger (1980)) J_m^d in the space of functions $\mathbb{R}^d \to \mathbb{R}$ where it can be finitely computed as:

$$J_m^d(\eta, \zeta) := \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1! \cdots \alpha_d!} \int_{\mathbb{R}^d} \left(\frac{\partial^m \eta}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \right) \left(\frac{\partial^m \zeta}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \right) dx_1 \cdots dx_d. \quad (2.23)$$

The objective is to estimate η by minimizing the criterion (2.1) for $\mathbb{X} = (-\infty, \infty)^d$ subject to the penalty $J(\eta) = J_m^d(\eta, \eta) = J_m^d(\eta)$ which can be written as

$$J_m^d(\eta) := \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1! \cdots \alpha_d!} \int_{\mathbb{R}^d} \left(\frac{\partial^m \eta}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \right)^2 dx_1 \cdots dx_d, \quad (2.24)$$

and $\mathcal{H} = \left\{ \eta : \mathbb{X} \to \mathbb{R} | J_m^d(\eta) < \infty \right\}$. For η to be in $\mathcal{H}$ we have to be able to compute $J_m^d(\eta)$. By Lemma (67), the null space $N_{J_m^d} := \left\{ f \in H : J_m^d(f) = 0 \right\}$ of J_m^d consists of polynomials in d variables of order up to $m-1$ and is of finite dimension $l = \binom{d+m-1}{d}$. In order for $[x] : \mathcal{H} \to \mathbb{R}$ and $x \in \mathbb{X}$ defined as $[x]f := f(x)$ to be continuous, we need that $2m > d$ in $\mathcal{H}$ Duchon (1977); Meinguet (1979), and thus $\mathcal{H}$ is a *RKHS* . In this context, J_m^d is a square semi norm (Lemma (68)) and hence $\mathcal{H} \ominus N_{J_m^d}$ is a *RKHS* . Another property of the thin plate splines is that they are invariant to translations, rotations and contractions; interpolation on the contracted set $\{\lambda x : x \in A, A \subset \mathbb{X}\}$ is equivalent to interpolating in A and then applying the transformation $x \to \lambda x$ (Lemma (69)).

Example 1

When the domain $\mathbb{X} = \mathbb{R}$ $(d = 1)$*, the minimization criterion* (1.1) *with penalty* (2.24) *corresponds to the minimization of*

$$\sum_{i=1}^{n} (y_i - \eta(X_i))^2 + n\lambda \int_{-\infty}^{\infty} \left\{ \eta^{(m)}(x) \right\}^2 \mathrm{d}x.$$

For $m = 1$ *the penalty corresponds to linear splines, and when* $m = 2$*, we obtain the cubic smoothing splines (Reinsch (1967); Craven and Wahba (1978); Ruppert et al. (2003)). It is possible to penalize any derivative as long as* $2m > d$*.*

The existence and uniqueness of the minimum of the functional (2.1) for the thin plate splines follows from Propositions (48), (49) and Theorem 50. By Proposition (48) we have that $\sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i))^2$ is a continuous and convex functional in $\mathcal{H}$; J_m^d is a square semi-norm with finite null space; $\sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i))^2$ has a unique minimizer in $\mathcal{N}_{J_m^d}$, the least squares estimator when η is a polynomial in d covariates with degree $\leq m - 1$. By Theorem 50, the functional (2.1) has a minimizer in $\mathcal{H}$. By Proposition (48), if S is of full column rank and $\lambda > 0$ then (2.1) is strictly convex, and using Proposition (49) we conclude that (2.1) has a unique minimizer in the context of the thin plate splines.

One of the versions of the Representer Theorem (Theorem 52; Kimeldorf and Wahba (1971); Wahba and Wendelberger (1980); Schölkopf et al. (2001)), states that η_λ has the expression (2.3) where $\{\psi_i\}_{i=1}^{l}$ is a basis for $N_{J_m^d}$.

Example 2 *An example of a basis for* $N_{J_m^d}$ *with* $m = 3$ *and* $d = 2$ *is the set of polynomials*

$$\begin{aligned} &\psi_1(x_1, x_2) = 1, && \psi_3(x_1, x_2) = x_2 && \psi_5(x_1, x_2) = x_1^2, \\ &\psi_2(x_1, x_2) = x_1, && \psi_4(x_1, x_2) = x_1 x_2, && \psi_6(x_1, x_2) = x_2^2, \end{aligned}$$

where $x_1, x_2 \in \mathbb{R}$*.*

In the setting of Theorem 52, $R_{J_m^d}$ is the reproducing kernel of $\mathcal{H} \ominus N_{J_m^d}$. In the context of Section 2, $\{c_i\}_{i=1}^{n} \subset \mathbb{R}$ are chosen so that $S^\intercal \mathbf{c} = 0$ and S is the matrix with $(j, i)th$ entry $\psi_i(\mathbf{x}_j)$.

Two limitations of thin plate splines are mentioned by Barry et al. (1986). First, the Bayes fitting method with such splines involves a complicated prior covariance function. Second, the

use of a single smoothing parameter λ suggests that the function under estimation is equally smooth in every direction of the domain, an unrealistic assumption. Wahba (1981) suggest a solution that consists in letting the roughness penalty for η to depend on two covariates:

$$J(\eta) = \int_{\mathbb{R}^2} \left[\left(\frac{\partial^2 \eta}{\partial x_1^2} \right)^2 + 2\theta \left(\frac{\partial^2 \eta}{\partial x_1 \partial x_2} \right)^2 + \theta^2 \left(\frac{\partial^2 \eta}{\partial x_2^2} \right)^2 \right] dx_1 dx_2. \tag{2.25}$$

Unfortunately, the penalty (2.25) is no longer invariant to rotation.

2.1.1.1 Reproducing kernel for thin plate splines

The Representer Theorem (Theorem 52) states that the solution to the minimization problem (1.1) is a linear combination of a basis of the null space $N_{J_m^d}$ and functions defined in terms of the reproducing kernel $R_{J_m^d}$ and the sampling points $\{\mathbf{x}_i\}_{i=1}^n$. The reproducing kernel is also used in the context of tensor thin plate splines (Section 2.1.2). We use the reproducing kernel in Section 2.1.3 to help us express (2.4) as a linear mixed model. Finally, the reproducing kernel appears again in the posterior distributions that arise when we estimate the coefficients in the thin plate splines within a Bayesian framework in Section 3. In this Section, we give an expression for the reproducing kernel.

First we present the radial basis functions introduced by Wahba and Wendelberger (1980). Denote the euclidean norm as $\|\cdot\|$ and

$$E_{m,d}(r) = \begin{cases} \theta_{m,d} r^{2m-d} \log(r), & \text{d even, for } \theta_{m,d} = \frac{(-1)^{d/2+m+1}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!}, \\ \theta_{m,d} r^{2m-d}, & \text{d odd, for } \theta_{m,d} = \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} \end{cases} \tag{2.26}$$

so that for a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, a radial basis is $\{E_{m,d}\|\cdot - \mathbf{X}_i\|\}_{i=1}^n$.

Second we need a reproducing kernel in the null space of J, $\mathcal{N}_{J_m^d} := \left\{\eta \in \mathcal{H} | J_m^d(\eta) = 0\right\}$ which is a complete linear space by Theorem 58. In order to describe the structure of *RKHS* for $\mathcal{N}_{J_m^d}$, we need an inner product: for some $N \in \mathbb{N}$, $\{u_i\}_{i=1}^N \subset \mathbb{R}^d$, $\{p_i\}_{i=1}^N \subset \mathbb{R}$ with $p_i > 0$ and $\sum_{i=1}^N p_i = 1$ define

$$(\eta, \zeta)_0 = \sum_{i=1}^N p_i \eta(u_i).\zeta(u_i). \tag{2.27}$$

By Lemma 67 and Proposition 59, $\{u_i\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$ are such that the matrix with $(i,j)th$ entry $(\psi_i, \psi_j)_0$ is non-singular, where $\{\psi_i\}_{i=1}^j$ a fixed basis of $\mathcal{N}_{J_m^d}$. These assumptions are sufficient for $\mathcal{N}_{J_m^d}$ to be a *RKHS* with (2.27) the inner product in $\mathcal{N}_{J_m^d}$.

In order to find and compute the reproducing kernel R_0 of $\mathcal{N}_{J_m^d}$ with inner product (2.27), we need an orthonormal basis $\{\phi_i\}_{i=1}^l \subset \mathcal{N}_{J_m^d}$ with $\phi_1(\mathbf{x}) = 1$. By the Gram-Schimdt normalization (Hoffman and Kunze (1990); Golub and Van Loan (2012)), given $\{\psi_i\}_{i=1}^l$ a set of polynomials that span $\mathcal{N}_{J_m^d}$, we can transform them and find such orthonormal basis. Explicit expressions for and an example of orthonormal basis $\{\phi_i\}_{i=1}^l$ are given in Proposition 62. The reproducing kernel in $\mathcal{N}_{J_m^d}$ by Proposition 59 is then

$$R_0(x,y) = \sum_{i=1}^{l} \phi_i(x)\phi_i(y). \tag{2.28}$$

Given the inner product (2.27) (or the choices of $\{u_i\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$), any orthonormal basis will lead to the same R_0 by Theorem 56, which states that the reproducing kernel is unique provided it exists.

We now present the projection of $f \in \mathcal{H}$ onto $\mathcal{N}_J$ since we need it for the reproducing kernel of $\mathcal{H} \ominus \mathcal{N}_{J_m^d}$. Let $\eta \in \mathcal{H} = \mathcal{N}_{J_m^d} \bigoplus \left(\mathcal{H} \ominus \mathcal{N}_{J_m^d}\right)$ or equivalently $\eta = \eta_0 + \eta_1$ with $\eta_0 \in \mathcal{N}_J$ and $\eta_1 \in \mathcal{H} \ominus \mathcal{N}_{J_m^d}$ for unique η_0 and η_1 (existence and uniqueness is ensured by Theorem 47), the projection of η onto $\mathcal{N}_J$ is by definition $P\eta = \eta_0$. By Proposition 60, the projection of $\eta \in \mathcal{H}$ onto is $\mathcal{N}_{J_m^d}$ is

$$(Pf)(\mathbf{x}) = \sum_{\nu=1}^{l} (f, \phi_\nu)_0 \, \phi_\nu(\mathbf{x}). \tag{2.29}$$

Define now the bi-linear form R_1 as

$$R_1(\mathbf{x},\mathbf{y}) = \left(I - P_{(\mathbf{x})}\right)\left(I - P_{(\mathbf{y})}\right) E\left(\|\mathbf{x}-\mathbf{y}\|\right), \tag{2.30}$$

where I is the identity operator and $P_{(x)}$ and $P_{(y)}$ are the projection operators defined by applying (2.29) to the arguments x and y, while E is given by (2.26). By Proposition 61, R_1 is the reproducing kernel of $\mathcal{H} \ominus \mathcal{N}_J$ with inner product J_m^d. R_1 is symmetric by the properties of $\|\cdot\|$ and the projections. To show that R_1 is non-negative definite we need to show that $J_m^d\left(R_1(x,\cdot), R_1(y,\cdot)\right) = R_1(x,y)$. The reproducing property $J_m^d\left((I-P)f, R_1(x,\cdot)\right) =$

$(I - P)f(x)$ for $f \in \mathcal{H}$ is more challenging to demonstrate; the idea of the proof is given in Proposition 61.

While it seems at first sight that R_1 depends on the choice of $\{u_i\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$ this is not the case. Note that (2.30) is a function of $\{u_i\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$ only through the projection (2.29). By Theorem 47 the projection onto $\mathcal{N}_{J_m^d}$ is unique and in consequence the operator P is invariant to the representation of (2.27). For computational reasons we set $N = n$, $u_i = \mathbf{x}_i$ and $p_i = 1/n$ as it is explained in Section 2.1.1.2.

2.1.1.2 Computation of reproducing kernel R_1 for thin plate splines

In this Section, we derive an expression for the $L_1 \times L_2$ matrix with $(i,j)th$ entry $R_1(\mathbf{x}_i, \mathbf{y}_j)$ for $\{\mathbf{x}_i\}_{i=1}^{L_1}, \{\mathbf{z}_i\}_{i=1}^{L_2} \subset \mathbb{X}$. Let $\mathbf{x}, \mathbf{z} \in \mathbb{X}$ and $\{u_i\}_{i=1}^N \subset \mathbb{X}$, $\{p_i\}_{i=1}^N \subset \mathbb{R}$ with properties listed in Section 2.1.1.1. We first expand R_1:

$$
\begin{aligned}
R_1(\mathbf{x}, \mathbf{z}) &= \left(I - P_{(\mathbf{x})}\right)\left(I - P_{(\mathbf{z})}\right) E\left(\|\mathbf{x} - \mathbf{z}\|\right) \\
&= \left(I - P_{(\mathbf{x})}\right)\left\{ E\left(\|\mathbf{x} - \mathbf{z}\|\right) - \sum_{\nu=1}^{l}\sum_{i=1}^{N} p_i E\left(\|\mathbf{x} - \mathbf{u}_i\|\right)\phi_\nu(\mathbf{u}_i)\phi_\nu(\mathbf{z}) \right\} \\
&= E\left(\|\mathbf{x} - \mathbf{z}\|\right) && (2.31) \\
&\quad - \sum_{\nu=1}^{l}\sum_{i=1}^{N} p_i E\left(\|\mathbf{x} - \mathbf{u}_i\|\right)\phi_\nu(\mathbf{u}_i)\phi_\nu(\mathbf{z}) && (2.32) \\
&\quad - \sum_{\nu=1}^{l}\sum_{i=1}^{N} p_i E\left(\|\mathbf{z} - \mathbf{u}_i\|\right)\phi_\nu(\mathbf{u}_i)\phi_\nu(\mathbf{x}) && (2.33) \\
&\quad + \sum_{\nu=1}^{l}\sum_{\mu=1}^{l} \phi_\nu(\mathbf{x})\phi_\mu(\mathbf{z}) \sum_{i=1}^{N}\sum_{j=1}^{N} p_i p_j \phi_\nu(\mathbf{u}_i)\phi_\mu(\mathbf{u}_j) E\left(\|\mathbf{u}_i - \mathbf{u}_j\|\right). && (2.34)
\end{aligned}
$$

The term (2.32) can be written as

$$
- \begin{pmatrix} E\left(\|\mathbf{x} - \mathbf{u}_1\|\right) \\ E\left(\|\mathbf{x} - \mathbf{u}_2\|\right) \\ \vdots \\ E\left(\|\mathbf{x} - \mathbf{u}_N\|\right) \end{pmatrix}^{\intercal} \begin{pmatrix} p_1\phi_1(\mathbf{u}_1) & \cdots & p_1\phi_l(\mathbf{u}_1) \\ p_2\phi_1(\mathbf{u}_2) & \cdots & p_2\phi_l(\mathbf{u}_2) \\ \vdots & \ddots & \vdots \\ p_N\phi_1(\mathbf{u}_N) & \cdots & p_N\phi_l(\mathbf{u}_N) \end{pmatrix} \begin{pmatrix} \phi_1(\mathbf{z}) \\ \phi_2(\mathbf{z}) \\ \vdots \\ \phi_l(\mathbf{z}) \end{pmatrix}, \qquad (2.35)
$$

the term (2.33) can be written as

$$-\begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \vdots \\ \phi_l(\mathbf{x}) \end{pmatrix}^{\intercal} \begin{pmatrix} p_1\phi_1(\mathbf{u}_1) & \cdots & p_N\phi_1(\mathbf{u}_N) \\ p_1\phi_2(\mathbf{u}_1) & \cdots & p_N\phi_2(\mathbf{u}_N) \\ \vdots & \ddots & \vdots \\ p_1\phi_l(\mathbf{u}_1) & \cdots & p_N\phi_l(\mathbf{u}_N) \end{pmatrix} \begin{pmatrix} E\left(\|\mathbf{z}-\mathbf{u}_1\|\right) \\ E\left(\|\mathbf{z}-\mathbf{u}_2\|\right) \\ \vdots \\ E\left(\|\mathbf{z}-\mathbf{u}_N\|\right) \end{pmatrix}, \tag{2.36}$$

and the term (2.34) can be written as

$$\begin{pmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \vdots \\ \phi_l(\mathbf{x}) \end{pmatrix}^{\intercal} \begin{pmatrix} p_1\phi_1(\mathbf{u}_1) & \cdots & p_N\phi_1(\mathbf{u}_N) \\ p_1\phi_2(\mathbf{u}_1) & \cdots & p_N\phi_2(\mathbf{u}_N) \\ \vdots & \ddots & \vdots \\ p_1\phi_l(\mathbf{u}_1) & \cdots & p_N\phi_l(\mathbf{u}_N) \end{pmatrix} \begin{pmatrix} E\left(\|\mathbf{u}_1-\mathbf{u}_1\|\right) & \cdots & E\left(\|\mathbf{u}_1-\mathbf{u}_N\|\right) \\ E\left(\|\mathbf{u}_2-\mathbf{u}_1\|\right) & \cdots & E\left(\|\mathbf{u}_2-\mathbf{u}_N\|\right) \\ \vdots & \ddots & \vdots \\ E\left(\|\mathbf{u}_N-\mathbf{u}_1\|\right) & \cdots & E\left(\|\mathbf{u}_N-\mathbf{u}_N\|\right) \end{pmatrix} \begin{pmatrix} p_1\phi_1(\mathbf{u}_1) & \cdots & p_1\phi_l(\mathbf{u}_1) \\ p_2\phi_1(\mathbf{u}_2) & \cdots & p_2\phi_l(\mathbf{u}_2) \\ \vdots & \ddots & \vdots \\ p_N\phi_1(\mathbf{u}_N) & \cdots & p_N\phi_l(\mathbf{u}_N) \end{pmatrix} \begin{pmatrix} \phi_1(\mathbf{z}) \\ \phi_2(\mathbf{z}) \\ \vdots \\ \phi_l(\mathbf{z}) \end{pmatrix}. \tag{2.37}$$

Therefore, putting it all together, we can write the $L_1 \times L_2$ matrix with $(i,j)th$ entry $R_1(\mathbf{x}_i, \mathbf{z}_j)$ for $\mathbf{x}_i, \mathbf{z}_j \in \mathbb{X}$ as

$$\left(R_1\left(\mathbf{x}_i, \mathbf{z}_j\right)\right)_{\substack{i=1,\cdots,L_1 \\ j=1,\cdots,L_2}} = K_{xz} - K_{xu}A_1A_3^{\intercal} - A_2A_1^{\intercal}K_{uz} + A_2A_1^{\intercal}K_{uu}A_1A_3^{\intercal}, \tag{2.38}$$

where

$$K_{xz} = \begin{pmatrix} E\left(\|\mathbf{x}_1-\mathbf{z}_1\|\right) & E\left(\|\mathbf{x}_1-\mathbf{z}_2\|\right) & \cdots & E\left(\|\mathbf{x}_1-\mathbf{z}_{L_2}\|\right) \\ E\left(\|\mathbf{x}_2-\mathbf{z}_1\|\right) & E\left(\|\mathbf{x}_2-\mathbf{z}_2\|\right) & \cdots & E\left(\|\mathbf{x}_2-\mathbf{z}_{L_2}\|\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(\|\mathbf{x}_{L_1}-\mathbf{z}_1\|\right) & E\left(\|\mathbf{x}_{L_1}-\mathbf{z}_2\|\right) & \cdots & E\left(\|\mathbf{x}_{L_1}-\mathbf{z}_{L_2}\|\right) \end{pmatrix}$$

$$K_{xu} = \begin{pmatrix} E\left(\|\mathbf{x}_1-\mathbf{u}_1\|\right) & E\left(\|\mathbf{x}_1-\mathbf{u}_2\|\right) & \cdots & E\left(\|\mathbf{x}_1-\mathbf{u}_N\|\right) \\ E\left(\|\mathbf{x}_2-\mathbf{u}_1\|\right) & E\left(\|\mathbf{x}_2-\mathbf{u}_2\|\right) & \cdots & E\left(\|\mathbf{x}_2-\mathbf{u}_N\|\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(\|\mathbf{x}_{L_1}-\mathbf{u}_1\|\right) & E\left(\|\mathbf{x}_{L_1}-\mathbf{u}_2\|\right) & \cdots & E\left(\|\mathbf{x}_{L_1}-\mathbf{u}_N\|\right) \end{pmatrix}$$

$$K_{uz} = \begin{pmatrix} E\left(\|\mathbf{u}_1-\mathbf{z}_1\|\right) & E\left(\|\mathbf{u}_1-\mathbf{z}_2\|\right) & \cdots & E\left(\|\mathbf{u}_1-\mathbf{z}_{L_2}\|\right) \\ E\left(\|\mathbf{u}_2-\mathbf{z}_1\|\right) & E\left(\|\mathbf{u}_2-\mathbf{z}_2\|\right) & \cdots & E\left(\|\mathbf{u}_2-\mathbf{z}_{L_2}\|\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(\|\mathbf{u}_N-\mathbf{z}_1\|\right) & E\left(\|\mathbf{u}_N-\mathbf{z}_2\|\right) & \cdots & E\left(\|\mathbf{u}_N-\mathbf{z}_{L_2}\|\right) \end{pmatrix}$$

$$K_{uu} = \begin{pmatrix} E\left(\|\mathbf{u}_1 - \mathbf{u}_1\|\right) & E\left(\|\mathbf{u}_1 - \mathbf{u}_2\|\right) & \cdots & E\left(\|\mathbf{u}_1 - \mathbf{u}_N\|\right) \\ E\left(\|\mathbf{u}_2 - \mathbf{u}_1\|\right) & E\left(\|\mathbf{u}_2 - \mathbf{u}_2\|\right) & \cdots & E\left(\|\mathbf{u}_2 - \mathbf{u}_N\|\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(\|\mathbf{u}_N - \mathbf{u}_1\|\right) & E\left(\|\mathbf{u}_N - \mathbf{u}_2\|\right) & \cdots & E\left(\|\mathbf{u}_N - \mathbf{u}_N\|\right) \end{pmatrix}$$

$$A_1 = \begin{pmatrix} p_1\phi_1(\mathbf{u}_1) & p_1\phi_2(\mathbf{u}_1) & \cdots & p_1\phi_l(\mathbf{u}_1) \\ p_2\phi_1(\mathbf{u}_2) & p_2\phi_2(\mathbf{u}_2) & \cdots & p_2\phi_l(\mathbf{u}_2) \\ \vdots & \vdots & \ddots & \vdots \\ p_N\phi_1(\mathbf{u}_N) & p_N\phi_2(\mathbf{u}_N) & \cdots & p_N\phi_l(\mathbf{u}_N) \end{pmatrix}$$

$$A_2 = \begin{pmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_l(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_l(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_{L_1}) & \phi_2(\mathbf{x}_{L_1}) & \cdots & \phi_l(\mathbf{x}_{L_1}) \end{pmatrix}$$

$$A_3 = \begin{pmatrix} \phi_1(\mathbf{z}_1) & \phi_2(\mathbf{z}_1) & \cdots & \phi_l(\mathbf{z}_1) \\ \phi_1(\mathbf{z}_2) & \phi_2(\mathbf{z}_2) & \cdots & \phi_l(\mathbf{z}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{z}_{L_2}) & \phi_2(\mathbf{z}_{L_2}) & \cdots & \phi_l(\mathbf{z}_{L_2}) \end{pmatrix}.$$

The matrices A_1, A_2 and A_3 require that we evaluate the orthonormal (per (2.27)) basis $\{\phi_i\}_{i=1}^l$. One approach to compute the matrix $\mathbf{\Phi}_{i,j} = \phi_j(\mathbf{x}_i)$ is to use Proposition 62 which uses the matrix $\mathbf{\Psi}_{i,j} = \psi_j(\mathbf{x}_i)$, with $\{\psi_i\}_{i=1}^l$ any basis of $\mathcal{N}_{J_m^d}$. For $N = n$, $\mathbf{u}_i = \mathbf{x}_i$ and $p_i = 1/n$, Proposition 62 states that given the QR-decomposition of $\mathbf{\Psi} = (F_1\, F_2)\binom{R}{\mathbf{0}} = F_1 R$ we can write $\phi_i = \sqrt{n}\sum_{j=1}^l \left(R^{-\intercal}\right)_{i,j}\psi_j$ and in particular $\mathbf{\Phi} = \sqrt{n}F_1$.

2.1.2 Tensor product splines using thin plate splines as marginals

The idea of the tensor product smoothing splines was first described by Barry et al. (1986) and Wahba (1987) while other advances on the topic since then are in (Gu and Wahba (1993a,b); Chen (1993); Barry et al. (1986); Gu (2013)). The tensor product smoothing splines is a way to construct a *RKHS* of (in particular but not restricted to) real functions over the product space $\prod_{i=1}^{L} \mathbb{X}_i$, which is achieved through the tensor product of the *RKHS's* $\{\mathcal{H}_{(i)}\}_{i=1}^{L}$ of real functions, each of them over the domains $\{\mathbb{X}_i\}_{i=1}^{L}$. For example, the thin plate splines from Section 2.1.1 may be used to define each of the $\mathcal{H}_{(i)}$ as spaces of real functions of one covariate.

The construction of the tensor product of *RKHS's* follows from Theorems 56 and 63. Theorem 56 states that one may completely describe a *RKHS* over a domain $\mathbb{X}$ simply by specifying a non-negative definite function on $\mathbb{X} \times \mathbb{X}$; Theorem 63 provides the means to specify a non-negative definite function on $\prod_{i=1}^{L} \mathbb{X}_i$ as the product of the reproducing kernels $R(x, y) := \prod_{i=1}^{L} R_{(i)}(x_{(i)}, y_{(i)})$, where $x_{(i)} \in \mathbb{X}_i$ denotes the i^{th} coordinate of $x \in \mathbb{X}$.

The *RKHS* $\mathcal{H}$ corresponding to R is the tensor product space of $\mathcal{H}_{(1)}$, $\mathcal{H}_{(2)}, \cdots$ and $\mathcal{H}_{(L)}$ and is denoted as $\mathcal{H} = \bigotimes_{i=1}^{L} \mathcal{H}_{(i)}$.

If, furthermore, each of the reproducing kernels has a direct sum decomposition of the form $\mathcal{H}_{(i)} = \mathcal{H}_{0\,(i)} \bigoplus \mathcal{H}_{1\,(i)}$ where $\mathcal{H}_{0\,(i)} = \mathbb{R}$ has a reproducing kernel $R_{0\,(i)} \propto 1$ and $\mathcal{H}_{1\,(i)}$ has a reproducing kernel $R_{1\,(i)}$, an averaging operator A_i is required by identifiability reasons and the condition $A_i R_{1\,(i)}(x_{(i)}, \cdot) = 0 \; \forall x_{(i)} \in \mathbb{X}_i$ is needs to be satisfied (Gu (2013)). Then the tensor product space can be re written as

$$\mathcal{H} = \bigotimes_{i=1}^{L} \left(\mathcal{H}_{0\,(i)} \bigoplus \mathcal{H}_{1\,(i)} \right) = \bigoplus_{\mathcal{S} \in 2^{\{1,\ldots,L\}}} \left\{ \left(\bigotimes_{i \notin \mathcal{S}} \mathcal{H}_{0\,(i)} \right) \bigotimes \left(\bigotimes_{i \in \mathcal{S}} \mathcal{H}_{1\,(i)} \right) \right\} = \bigoplus_{\mathcal{S} \in 2^{\{1,\ldots,L\}}} \mathcal{H}_{\mathcal{S}}. \tag{2.39}$$

Each of the spaces $\mathcal{H}_{\mathcal{S}}$ has reproducing kernel $R_{\mathcal{S}} \propto \prod_{i \in \mathcal{S}} R_{1\,(i)}$ by Theorem 63, and the reproducing kernel of (2.39) by Theorem 56 would be of the form

$$R_J = \sum_{\mathcal{S} \in 2^{\{1,\ldots,L\}}} R_{\mathcal{S}}, \tag{2.40}$$

but allowing for re-scaling using $\{\theta_{\mathcal{S}}\} \in \mathbb{R}$ no negatives, we may take

$$R_K = \sum_{\mathcal{S} \in 2^{\{1,\ldots,L\}}} \theta_{\mathcal{S}} R_{\mathcal{S}},$$

which, by Theorems 56 and 63, R_K is a reproducing kernel in the *RKHS* (2.39) with its respective inner product K.

The minimizer of (2.1), allowing K to be a semi-inner product in the tensor product spline $\mathcal{H}$, is called *tensor product smoothing spline* and the marginals are the spaces $\mathcal{H}_{(i)}$. This is the setting for the minimization problem described with expressions (2.1) and penalty (2.2) with $p > 1$.

2.1.3 Linear mixed model as interpretation for the solution of the penalized least squares minimization problem

Robinson (1991) presents results for the estimation of the BLUP for linear mixed models in the frequentist setting. There are different ways to arrive at a BLUP in the mixed model context. Below, we describe the solution proposed by Henderson (Henderson (1950, 1973)) and discussed by Robinson. Consider the model

$$\mathbf{y}|\mathbf{d},\, \mathbf{c},\, \mathbf{e} = V\mathbf{d} + Z\mathbf{c} + \mathbf{e}, \tag{2.41}$$

where $\mathbf{y} \in \mathcal{M}_{n\times 1}(\mathbb{R})$ is column vector of observable responses, $\mathbf{d}$ is a vector of l unknown fixed parameters, $V \in \mathcal{M}_{n\times l}(\mathbb{R})$ is a known, full column rank matrix, and $Z \in \mathcal{M}_{n\times n}(\mathbb{R})$ is a known, symmetric matrix. The vectors $\mathbf{c} \in \mathcal{M}_{n\times 1}(\mathbb{R})$, $\mathbf{e} \in \mathcal{M}_{n\times 1}(\mathbb{R})$ are non observable random variables such that

$$\begin{pmatrix} \mathbf{c} \\ \mathbf{e} \end{pmatrix} \sim \mathbf{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_c^2\mathbf{G} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2\mathbf{R} \end{pmatrix}\right) \tag{2.42}$$

where $\mathbf{G} \in \mathcal{M}_{n\times n}(\mathbb{R})$ and $\mathbf{R} \in \mathcal{M}_{n\times n}(\mathbb{R})$ are known positive definite matrices, and $\sigma_e > 0$ and $\sigma_c > 0$ are fixed parameters which may be unknown but for which we can obtain unbiased estimates $\hat{\sigma_c^2}$ and $\hat{\sigma_e^2}$ ($\hat{\sigma_c^2}$, and $\hat{\sigma_e^2}$ are not needed for this thesis). A well-known result is that

$$\begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \end{pmatrix} = \underset{\mathbf{d},\mathbf{c}}{\arg\min} \left(\mathbf{y} - V\mathbf{d} + Z\mathbf{c}\right)^{\intercal} \mathbf{R}^{-1} \left(\mathbf{y} - V\mathbf{d} + Z\mathbf{c}\right) + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{c}^{\intercal}\mathbf{G}^{-1}\mathbf{c}, \tag{2.43}$$

where the BLUP solution can be written as

$$\begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \end{pmatrix} = \left(C^{\intercal}\mathbf{R}^{-1}C + B\right)^{-1} C^{\intercal}\mathbf{R}^{-1}\mathbf{y}. \tag{2.44}$$

Here, $C = [V\ Z]$ is the matrix formed by stacking together the columns of V and Z and

$$B = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_c^2}\mathbf{G}^{-1} \end{pmatrix}.$$

Note that expression (2.4) from the penalized least square minimization is similar to (2.43) if we choose $V = S$ and $Z = Q$ from Section 2, $\mathbf{G} = Q^+$ (definition 32), $\mathbf{R} = I_n$ and $n\lambda = \frac{\sigma_e^2}{\sigma_c^2}$. In this case, we see by (2.44) that the following expression has to be invertible:

$$C^\intercal \mathbf{R}^{-1} C + B = \begin{pmatrix} S^\intercal S & S^\intercal Q \\ QS & QQ + \frac{\sigma_e^2}{\sigma_c^2} Q \end{pmatrix}. \tag{2.45}$$

Expression (2.45) is not invertible in our context because Q is not necessarily invertible: if (2.45) was invertible we would have that (2.5) has a unique solution, but we have established that it does not have a unique solution, at least in the case of the thin plate splines and tensor thin plate splines.

To get around the fact that (2.45) is not invertible, consider a radial basis $\{E_{m,d}\|\cdot - \mathbf{x}_i\|\}_{i=1}^n$ from (2.26) instead of the full expression (2.30). The matrix Z now has $Z_{i,j} = E_{m,d}\|\mathbf{x}_i - \mathbf{x}_j\|$, but $V = S$ as in the thin plate splines, where $S_{ij} = \psi_j(\mathbf{x}_i)$, ψ_i an element of a fixed basis for the polynomials in d covariates and degree smaller than m. Ruppert et al. (2003) uses this model as a first step to estimate η in (1.1) which implies that J is no longer (2.23).

Even if the inverse of

$$C^\intercal \mathbf{R}^{-1} C + B = \begin{pmatrix} S^\intercal S & S^\intercal Z \\ ZS & ZZ + \frac{\sigma_e^2}{\sigma_c^2} Z \end{pmatrix}$$

exists, the choice $\mathbf{G} = Z^{-1}$ expressions (2.41), (2.42) do not identify a proper model because $Cov(Z\mathbf{c}) = \sigma_c^2 Z$ and Z is symmetric but not positive definite. An approximation to the solution can be obtained by changing the co-variance of $\mathbf{c}$ to $Cov(\mathbf{c}) = \sigma_c^2 Z^{-\frac{1}{2}} \left(Z^{-\frac{1}{2}}\right)^\intercal$, where $Z^{-\frac{1}{2}}$ is the inverse square root matrix of Z defined using its singular value decomposition (definition 33); the co-variance of $\mathbf{c}$ is represented in this way with a positive definite matrix if Z is full column rank, and with a semi positive definite matrix if Z is not full column rank.

The final model to approximate the solution of (1.1) through (2.4) with a linear mixed model can be achieved with the change of variables $Z^\star = ZZ^{-\frac{1}{2}} = Z^{\frac{1}{2}}$, so that

$$\begin{aligned} \mathbf{y}|\mathbf{c},\, \mathbf{e} &= S\mathbf{d} + Z^{\star}\mathbf{c} + \mathbf{e}, \\ \begin{pmatrix} \mathbf{c} \\ \mathbf{e} \end{pmatrix} &\sim \mathbf{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_c^2 I_k & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 I_n \end{pmatrix}\right), \end{aligned} \tag{2.46}$$

and where (using (2.44)):

$$\begin{pmatrix} \hat{\mathbf{d}^{\star}} \\ \hat{\mathbf{c}^{\star}} \end{pmatrix} = \left\{ [S\ Z^{\star}]^{\intercal}\,[S\ Z^{\star}] + \begin{pmatrix} \mathbf{0}_{l\times l} & \mathbf{0}_{l\times n} \\ \mathbf{0}_{n\times l} & \frac{\sigma_e^2}{\sigma_c^2}\mathbf{I}_{n\times n} \end{pmatrix} \right\}^{-1} [S\ Z^{\star}]^{\intercal}\,\mathbf{y}. \tag{2.47}$$

The advantage of this interpretation of a solution to the penalized least square (1.1) through (2.3) is that the smoothing parameter λ can be estimated as the ratio of two variance components if $\{\theta_i\}_{i=1}^{p}$ are given, estimations are provided or $p = 1$ and the only smoothing parameter is λ. The variance components can be estimated using different approaches and we revisit this topic in 3.2.3. In 2.3 we discuss three different approaches to select λ and $\{\theta_i\}_{i=1}^{p}$.

2.2 Efficient Approximated Solution to the Non-Parametric Regression Problem

Section 2.2.1 describes the approximation to the solution of the penalized least square minimization problem (1.1) proposed by Gu and Kim (2002) and expanded in Kim and Gu (2004); in Section 2.2.2 we explain an adaptation of the Linear Mixed Model interpretation from Section 2.1.3.

2.2.1 Smoothing splines

Kim and Gu 2004 approach the problem of penalized likelihood regression in a general way that includes a variety of link functions for exponential families. They propose an approximated solution to the regression problem, which lies in a lower dimensional space of functions. They show that the approximated solution has the same asymptotic convergence rate as the exact solution, but computations are of order $O(kn^2)$, $k << n$ rather than order $O(n^3)$. Hence, this approximation allows the implementation of much faster numerical algorithms. For the approximation, it is assumed that a set of knots $\{\mathbf{z}_i\}_{i=1}^{k} \subset \mathbf{X} := \{\mathbf{x}_i\}_{i=1}^{n} \subset \mathbb{X}$ with the same

limiting density as $\mathbf{X}$ is given. A small number of knots k is preferred for computational efficiency but k too small may decrease the statistical performance. Section 2.2.3 includes an empirical rule for the selection of k and a discussion of the algorithm that we adopt here to choose the knots.

As explained in Section 2, the solution to the penalized least square regression (1.1) is of the form (2.3). Gu and Kim (2002) approximate the solution by minimizing (2.1) in the space

$$\mathcal{H}^\star = \mathcal{N}_J \bigoplus span\{R_J(\mathbf{z}_i, \cdot), i = 1, ..., k\}. \tag{2.48}$$

Analogously to (2.3), any functions $\eta \in \mathcal{H}^\star$ can be written as

$$\begin{aligned} \eta(\mathbf{x}) &= \sum_{i=1}^{l} d_i \psi_i(\mathbf{x}) + \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \mathbf{x}) \\ &= \psi(\mathbf{x})^\intercal \mathbf{d} + \xi^\star(\mathbf{x})^\intercal \mathbf{c} \end{aligned} \tag{2.49}$$

with $\{\psi_\nu\}_{\nu=1}^{l}$ being a basis of $\mathcal{N}_J$ and

$$\begin{aligned} \mathbf{d} &= (d_1 \cdots d_l)^\intercal, d_i \in \mathbb{R} \\ \mathbf{c} &= (c_1 \cdots c_k)^\intercal, c_i \in \mathbb{R} \\ \psi(\mathbf{x}) &= (\psi_1(\mathbf{x}) \cdots \psi_l(\mathbf{x}))^\intercal \\ \xi^\star(\mathbf{x}) &= (R_J(\mathbf{z}_1, \mathbf{x}) \cdots R_J(\mathbf{z}_k, \mathbf{x}))^\intercal . \end{aligned}$$

A similar expression to (2.4) can be obtained as well by plugging in (2.49) to (2.1) and using the reproducing kernel property as stated in Proposition 53; the subsequent expression (2.50) is obtained as

$$(\mathbf{y} - S\mathbf{d} - R\mathbf{c})^\intercal (\mathbf{y} - S\mathbf{d} - R\mathbf{c}) + n\lambda \mathbf{c}^\intercal Q \mathbf{c}, \tag{2.50}$$

where $R \in \mathcal{M}_{n\times k}(\mathbb{R})$ with $(i,j)th$ entry $R_J(\mathbf{x}_i, \mathbf{z}_j)$, $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ with $(i,j)th$ entry $R_J(\mathbf{z}_i, \mathbf{z}_j)$ and $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ with $(i,j)th$ entry $\psi_j(x_i)$. Finally, similarly to (2.5) using now Lemma 64, the solutions to the linear system

$$\begin{pmatrix} S^\intercal S & S^\intercal R \\ R^\intercal S & R^\intercal R + n\lambda Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^\intercal \mathbf{y} \\ R^\intercal \mathbf{y} \end{pmatrix} \tag{2.51}$$

provide the function in $\mathcal{H}^\star$ that minimizes (2.1). The solutions to (2.51) permit finding the inflection points of (2.1) in $\mathcal{H}^\star$ but in order to prove that some of these are minimums we need

to assume that S is of full column rank as explained next. Observe that $span\{R_J(\mathbf{z}_i)\}_{i=1}^k$ is a closed subspace of $\mathcal{N}_J \ominus \mathcal{H}$ and furthermore it is a Hilbert space with the same inner product J and same reproducing kernel R_1 as $\mathcal{N}_J \ominus \mathcal{H}$. By Proposition 48, the functional $\sum_{i=1}^n (y_i - \eta(\mathbf{x}_i))^2$ is continuous and convex in $\mathcal{H}^\star$ and when S is of full column rank, the convexity is strict in $\mathcal{N}_J$ and the functional then has a minimizer in this space. Theorem 50 states that a solution to (1.1) exists in $\mathcal{H}^\star$ as long as $\sum_{i=1}^n (y_i - \eta(\mathbf{x}_i))^2$ has a minimizer in $\mathcal{N}_J$. Proposition 48 states that (2.1) is strictly convex in $\mathcal{H}^\star$ when S is of full column rank and by Proposition 49, (2.1) has a unique minimizer in $\mathcal{H}^\star$. Then if S is full column rank, a solution to (2.51) will lead to the unique solution to (1.1) in $\mathcal{H}^\star$ through (2.49). Even when there were multiple solutions to (2.51), they yield the same $\eta \in \mathcal{H}^\star$. In practice, a solution to (2.51) is chosen as described in Section 3 (Theorems 5 and 6); it will appear in a Bayesian setting as the mean of a posterior distribution.

2.2.2 Linear mixed model interpretation as an approximate solution to the penalized least squares minimization problem

We adopt the same notation as in Section 2.2.1. Here, we want to use form (2.49) of η, and taking advantage of the fact that the minimizing expression (2.50) leads to a point estimate of the solution of (1.1) in the space $\mathcal{H}^\star$, propose a linear mixed model interpretation as in the Section 2.1.3.

Following the arguments of Section 2.1.3, we can minimize an expression similar to (2.50):

$$\begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \end{pmatrix} = \arg\min_{\mathbf{d},\mathbf{c}} (\mathbf{y} - S\mathbf{d} + U\mathbf{c})^\intercal (\mathbf{y} - S\mathbf{d} + U\mathbf{c}) + \frac{\sigma_e^2}{\sigma_c^2} \mathbf{c}^\intercal V \mathbf{c}, \tag{2.52}$$

for $S_{i,j} = \phi_j(\mathbf{x}_i)$ (based on basis $\{\phi_i\}_{i=1}^l$ of $\mathcal{N}_{J_m^d}$), $U_{i,j} = E_{m,d}\|\mathbf{x}_i - \mathbf{z}_j\|$, and $V_{i,j} = E_{m,d}\|\mathbf{z}_i - \mathbf{z}_j\|$. Then, for the response model given by

$$\mathbf{y}|\mathbf{d},\, \mathbf{c},\, \mathbf{e} = S\mathbf{d} + U\mathbf{c} + \mathbf{e}$$

$$\begin{pmatrix} \mathbf{c} \\ \mathbf{e} \end{pmatrix} \sim \mathbf{N}_{n+k} \left(\mathbf{0}, \begin{pmatrix} \sigma_c^2 V^{-1/2} \left(V^{-1/2}\right)^\intercal & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 I_{n\times n} \end{pmatrix} \right),$$

the BLUP solution to (2.52) is given by

$$\begin{pmatrix} \hat{\mathbf{d}} \\ \hat{\mathbf{c}} \end{pmatrix} = \left\{ [S\ U]^\intercal [S\ U] + \frac{\sigma_e^2}{\sigma_c^2} \begin{pmatrix} \mathbf{0}_{l\times l} & \mathbf{0}_{l\times n} \\ \mathbf{0}_{n\times l} & V^{-1} \end{pmatrix} \right\}^{-1} [S\ U]^\intercal \mathbf{Y}. \tag{2.53}$$

as described in Section 2.1.3.

2.2.3 Knots for approximated solution

In Section 2.2, we assume that a set of knots $\{\mathbf{z}_i\}_{i=1}^k$ is given. In this section we describe how we choose the number k of knots as well as their location. We adopt one of the methods that have been proposed in the literature.

According to Ruppert et al. (2003) and Ruppert (2012), in penalized regression, the smoothing is mainly controlled by the penalty term (2.2) and the smoothing parameters λ and $\{\theta_i\}_{i=1}^k$, so that number of knots is not a crucial parameter. A good choice for k is such that k is sufficiently large relative to the sample size, but not so large to require excessive computation time. Kim and Gu (2004) suggest an approach to select k based on the convergence rates of the approximated and exact solutions to the minimization problem. These rates depend on the smoothness of the link function η to be estimated. Kim and Gu (2004) consider random subsets $\{\mathbf{z}_i\}_{i=1}^k$ of $\{\mathbf{x}_i\}_{i=1}^n$ and show that with $k \approx n^{\frac{2}{pr+1}+\epsilon}$ for some $r > 1$, $p \in [1, 2]$ and for any $\epsilon > 0$, the convergence rates of the two solutions are the same. They conclude, via simulation, that in practice the choice $k = an^{\frac{2}{9}}$ for $a \in \{5, ..., 15\}$, $a = 7, 8$ may suffice. For the simulation study we describe in Section 3.3, we use $k = \lfloor \max\left\{30, 10n^{2/9}\right\} \rfloor$.

We hypothesize that instead of taking random subsets $\{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n$, the knots could be chosen to mimic the distribution of the sample $\{\mathbf{x}_i\}_{i=1}^n$. *For example in the univariate case* $d = 1$, $\{\mathbf{z}_i\}_{i=1}^k$ *may be chosen to be* k *equally spaced empirical quantiles of* $\{\mathbf{x}_i\}_{i=1}^n$. In this case, we may have that the approximated solution from Section 2.2, even when the asymptotic convergence rate from Kim and Gu (2004) does not change.

In addition to Kim and Gu (2004), several other authors have proposed algorithms to select knots. Examples are (Royle and Nychka (1998); Wood (2003); Ruppert (2012); Spiriti et al. (2013)). We use the algorithm in (Royle and Nychka (1998)) and implemented in the package

`fields` (Bia et al. (2014), Douglas Nychka et al. (2015)). The criterion for knot selection in (Royle and Nychka (1998)) is as follows. If $\mathfrak{X} = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ is the full sample set, given a finite candidate set of knots $\mathfrak{D} \subset \mathbb{R}^d$ with $\mathfrak{d}_k \subset \mathfrak{D}$ a subset of size k, define the function $\boldsymbol{\Delta} : \{\mathfrak{d}_k | \mathfrak{d}_k \subset \mathfrak{D}, \|\mathfrak{d}_k\| = k\} \to \mathbb{R}$ as

$$\boldsymbol{\Delta}(\mathfrak{X}, \mathfrak{D}; \mathfrak{d}_k) = \left(\sum_{\mathbf{x} \in \mathfrak{X}} \left(\sum_{\mathbf{z} \in \mathfrak{d}_k} \|\mathbf{x} - \mathbf{z}\|^{\alpha} \right)^{\frac{\beta}{\alpha}} \right)^{\frac{1}{\beta}},$$

where $\|\cdot\|$ is a norm in $\mathbb{R}^d$ and $\alpha < 0$ and $\beta > 0$; we take the euclidean norm and $-\alpha = \beta = 2$. The criterion $\boldsymbol{\Delta}$ measures a type of average of how well the design $\mathfrak{d}_k$ covers each of the points $\mathbf{x}_i \in \mathfrak{X}$. The knots may be chosen as $\{\mathbf{z}_i\}_{i=1}^k = \arg\min_{\mathfrak{d}_k \in \mathfrak{D}} \boldsymbol{\Delta}(\mathfrak{X}, \mathfrak{D}; \mathfrak{d}_k)$.

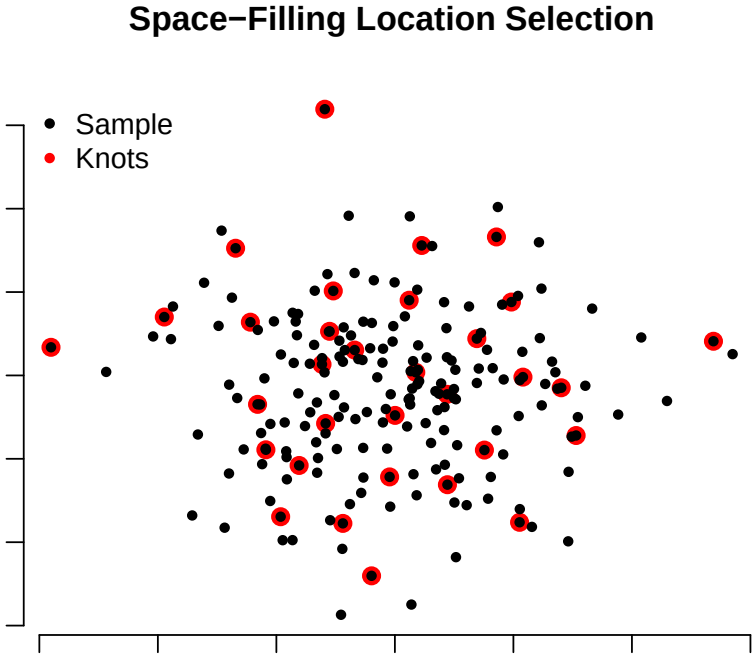

Figure 2.1 Example of the space filling location algorithm used to choose the knots from (Royle and Nychka (1998)). Observe that observations at the boundaries were chosen as knots.

In Chapter 3 and Section 5.2 we use this minimization scheme taking $\mathfrak{D} = \mathfrak{X}$. In Chapter 4 and Section 5.3, where we do not observe $\mathfrak{X} = \{\mathbf{x}_i\}_{i=1}^n$, but instead observe a contaminated labeled sample $\{(\mathbf{w}_{i1}, \cdots \mathbf{w}_{in_w})\}_{i=1}^n$, we take $\mathfrak{D} = \{\bar{\mathbf{w}}_{i\cdot}\}_{i=1}^n$. Figure 2.1 shows an example of the knots that are chosen with this algorithm using simulated points $\{\mathbf{x}_i\}_{i=1}^n$.

2.2.4 Degree of smoothness for the thin plate spline and tensor thin plate splines

The penalty J_m^d, (2.24), for the minimization problem (1.1) in the thin plate spline setting has a free parameter m, which controls the smoothness of the target function $\eta : \mathbb{R}^d \to \mathbb{R}$, e.g., for $d = 2$, the value $m = 2$ corresponds in some sense to the prior belief that

$$\int_{\mathbb{R}^2} \left[\left(\frac{\partial^2 \eta}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 \eta}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 \eta}{\partial x_2^2} \right)^2 \right] dx_1 dx_2,$$

is small; or in the case $d = 1$ and $m = 2$ correspond to the prior belief that $\int \eta^{(2)}(x)dx$ is small. We can determine an optimal value for m, using an algorithm that minimizes predictive mean squared error, and that can be implemented via generalized cross validation as in (Camber (1979)) or (Wahba and Wendelberger (1980)). Here, we just assume that the value m is known or, for the simulation studies in Sections 3.3 and 4.4, we consider a range of values for m.

2.3 Selection of Smoothing Parameters

The estimated function obtained through (2.3)- (2.4) with explicit representer given by (2.20), depends on smoothing parameters λ and $\theta_i's$. These parameters control the trade off between smoothness of the solution to (1.1) and how well the solution describes the data as measured by the quadratic loss function. Figure 2.2 shows three examples of the solution to (1.1) with different values of λ. A large value of λ leads to a solution close to a linear regression, while λ too small leads to a solution to (1.1) that interpolates the training set. Appropriate selection of the smoothing parameter is necessary.

In this section, we summarize four methods to approach the problem of choosing appropriate smoothing parameters. The first three methods were proposed in (Mallows (1973); Wahba and Craven (1978); Wahba (1985); Li (1986)). Such methods are based on the minimization of score functions that represent, for fixed λ and $\{\theta_i\}_{i=1}^p$, the performance (or approximation to the performance) of the estimator $\eta_{\lambda,\theta's}$ (η_λ for simplicity) obtained from a training set $\{\mathbf{x}_i, y_i\}_{i=1}^n$. The fourth approach is based on a linear mixed model interpretation of expression (2.4) as was described in Section 2.1.3.

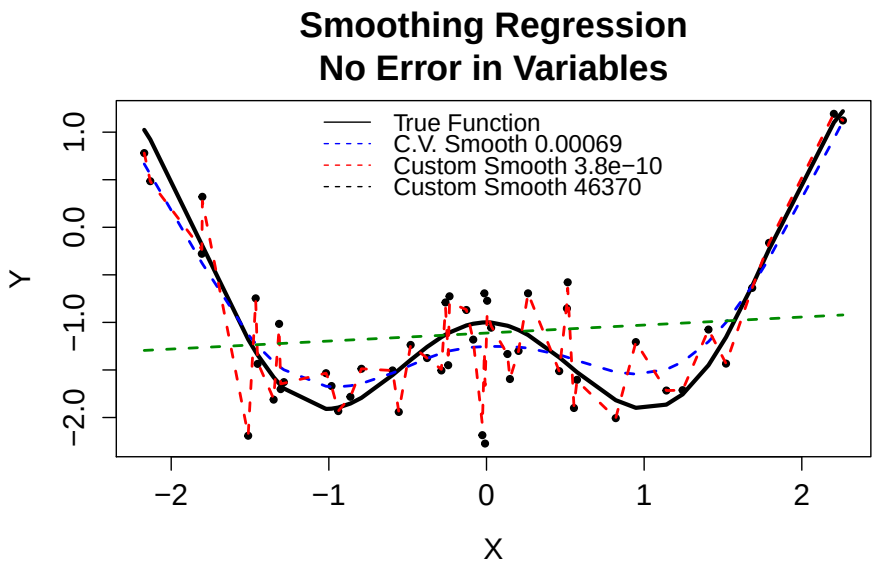

Figure 2.2 Basic example of the solution for the penalized least square minimization with different values of λ and $p = 1$, $\theta_1 = 1$. A value of λ too large provides a huge weight to the penalty term in (2.1) and decreases the effort of the solution to describe the data, resulting in a straight line which is the characteristic of an over smoothed solution. A value of λ too small increases the effort of the solution to the minimization problem to describe the data, leading to a solution that interpolates data points but at the cost of losing smoothness.

2.3.1 Unbiased estimate of relative loss

The content in this section appeared first in (Mallows (1973)) in the context of ridge regression. The performance of η_λ is assessed with the loss function $\mathfrak{L} : \mathbb{R}^+ \to \mathbb{R}$ defined as

$$\mathfrak{L}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\eta_\lambda(\mathbf{x}_i) - \eta(\mathbf{x}_i)\right)^2 . \tag{2.54}$$

By Theorem (50), the solution of (1.1) exists and is unique, hence there is not ambiguity in the definition of $\mathfrak{L}$; $\mathfrak{L}$ is a continuous function because the minimization of (2.1) is obtained through the solution of linear equations (2.51) which is a continuous process (small changes in the value of λ leads to small changes in the solution to (2.51)); $\mathfrak{L}$ can take on arbitrarily small values when λ is small; for λ fixed, $\sum_{i=1}^{n} \left(a_i - \eta(\mathbf{x}_i)\right)^2$, as a multivariate function of the vectors $(a_1, \cdots, a_n)$ is convex, hence there is a vector $(b_1, \cdots b_n)$ that minimizes it; therefore, by continuity of $\mathfrak{L}$, because $\mathfrak{L}$ can take values as close to 0 as desired, there must exist λ_0 (and θ_{i_0}'s) that minimizes (2.54) with $(\eta_{\lambda_0}(\mathbf{x}_1), \cdot, \eta_{\lambda_0}(\mathbf{x}_n)) = (b_1, \cdots b_n)$; more details in (Kurdila

and Zabarankin (2006)). A good choice for λ, given the observed data $(\mathbf{x}_i, y_i)_{i=1}^n$, would be the one that minimizes L.

Since we cannot compute (2.54) in practice, Mallows (1973) suggested approximating the relative loss $\mathfrak{L}(\lambda)$ with

$$\mathcal{U}(\lambda) = \frac{1}{n}\mathbf{y}^\intercal \left(I - A(\lambda)\right)^2 \mathbf{y} + 2\frac{\sigma^2}{n} trA(\lambda), \tag{2.55}$$

where $\mathbf{y} = (y_1\, y_2 \cdots y_n)^\intercal$, $\epsilon = (\epsilon_1\, \epsilon_2 \cdots \epsilon_n)^\intercal$ with $\mathbb{E}(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$ and $\{\epsilon_i\}_{i=1}^n$ are independent with bounded fourth moment. It can be shown (Gu, 2013, pag. 65) under the following mild condition

Condition 3 $\lim_{\substack{\lambda\to 0 \\ n\to\infty}} n\mathbb{E}\left(L(\lambda)\right) = \infty$

that (2.55) is an unbiased estimator of the relative loss $L + n^{-1}\epsilon^\intercal\epsilon$ in the sense

$$\mathcal{U}(\lambda) - L(\lambda) - n^{-1}\epsilon^\intercal\epsilon = o_p(L(\lambda)),$$

where o_p is defined in Definition 45. Since $n^{-1}\epsilon^\intercal\epsilon$ does not depend on λ, $U(\lambda)$ tracks $\mathfrak{L}(\lambda)$ closely. Observe that the minimizers λ_u of $U(\lambda)$ and λ_L of $\mathfrak{L}(\lambda)$ are stochastic. A formal justification for the approximation of $U(\lambda)$ to $L(\lambda)$ can be found in (Li (1986)).

Nevertheless, this method requires knowledge of the true value of σ^2. At least, we need a consistent estimator of σ^2 to plug into (2.55). Even with this disadvantage over other methods for choosing λ, U can be used in combination with a Bayesian approach as we describe in Chapter 3. The minimizer λ_u of (2.55) is called the *Unbiased Estimator of Relative Loss* (UERL).

2.3.2 Generalized cross validation

Cross validation techniques can be used as a strategy for choosing a good smoothing parameter λ using the observed realizations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. In this case, we want to minimize

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\eta_\lambda^{[i]}(\mathbf{x}_i) - y_i\right)^2$$

where $\eta_\lambda^{[l]}$ is the minimizer of the functional

$$\frac{1}{n}\sum_{i\neq l}\left(y_i - \eta(\mathbf{x}_i)\right)^2 + \lambda J(\eta). \tag{2.56}$$

As in (Gu (2013)), it is not necessary to solve (2.56) n times; instead, the delete-one operation can be carried out analytically and it can be shown that

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \eta_\lambda(\mathbf{x}_i))^2}{(1-a_{i,i}(\lambda))^2}, \tag{2.57}$$

where $a_{i,i}(\lambda)$ is the $(i,i)th$ entry of $A(\lambda)$ (equations (2.21) and (2.22)). Since typically, sampling points contribute unequally to the estimation of η, (Li (1986)) we can use a weighted version of V_0 such as

$$V_1(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\omega_i\frac{(y_i - \eta_\lambda(\mathbf{x}_i))^2}{(1-a_{i,i}(\lambda))^2}.$$

with weights $\{\omega_i\}_{i=1}^n$. If $\omega_i = n^2\frac{(1-a_{i,i}(\lambda))^2}{tr(I-A(\lambda))^2}$ is chosen, a generalized cross validation score is obtained (Wahba and Craven (1978); Li (1986); Gu (2013)) as

$$V(\lambda) = \frac{n^{-1}\mathbf{y}^\intercal\left(I-A(\lambda)\right)^2\mathbf{y}}{\left\{n^{-1}tr\left(I-A(\lambda)\right)\right\}^2}. \tag{2.58}$$

Under condition (3) and the following condition

Condition 4 $\lim_{\substack{\lambda\to 0\\ n\to\infty}}\frac{\{n^{-1}trA(\lambda)\}^2}{n^{-1}trA^2(\lambda)} = 0$

we can prove (Li, 1986, Proposition 3.1) that $V(\lambda)$ is a consistent estimator of the relative loss (2.54) such that $\mathfrak{L}(\lambda_v)/\mathfrak{L}(\lambda_0) = 1 + o(1)$, where λ_0 is the minimizer of (2.54) and λ_v is the minimizer of V. Even with the asymptotic behavior of (2.58), it is known (Kim and Gu (2004)) that the function V occasionally produces a minimizer resulting severe under-smoothing. A modified version with a fudge factor seems to be effective in preventing under-smoothing; where for a fudge factor $\alpha > 1$, we write

$$\mathcal{V}(\lambda) = \frac{n^{-1}\mathbf{y}^\intercal\left(I-A(\lambda)\right)^2\mathbf{y}}{\left\{n^{-1}tr\left(I-\alpha A(\lambda)\right)\right\}^2}. \tag{2.59}$$

Kim and Gu (2004) mention that an optimal value of α, if indeed there is an optimal one, would depend on the true function η and possibly other factors, but these are not available in practice. Through simulations was concluded that the empirical value $\alpha = 1.4$ provides adequate performance over a range of simulation settings. The minimizer λ_v of (2.59) is called the *Generalized Cross Validation* (GEV) estimator of λ.

Under the conditions (3) – (4) holding uniformly in a neighborhood of the optimal λ and as $\lambda \to 0$, $n \to \infty$, we have that $\mathfrak{L}(\lambda_v)/\mathfrak{L}(\lambda_0) = 1 + o(1)$, therefore λ_v and λ_0 are close to each

other and thus λ_v and λ_u are close to each other. Differentiating (2.55) and (2.58), setting the derivatives to zero and $\lambda_v = \lambda_u$, then equating expression of the derivatives set to zero and solving for σ^2, a consistent estimator of σ^2 is obtained as (Gu, 2013, Theorem 3.4):

$$\hat{\sigma}_v^2 = \frac{\mathbf{y}^\intercal \left(I - A(\lambda_v)\right)^2 \mathbf{y}}{tr\left(I - A(\lambda_v)\right)}. \tag{2.60}$$

2.3.3 Restricted maximum likelihood

The method described in this section is not designed to select smoothing parameters by minimizing any specific loss functions, but instead smoothing parameters arise in the context of some parameters in a Bayesian model based fitting approach. The method was first introduced by Wahba (Wahba (1985)) but related ideas were described previously by Wecker *et.al.* (Wecker and Ansley (1983)).

For observed $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, smoothing parameters λ and $\{\theta_i\}_{i=1}^p$ can be determined via Restricted Maximum Likelihood (RML) under the model

$$\begin{aligned} y_i &= \eta(\mathbf{x}_i) + \epsilon_i, \\ \epsilon_i &\overset{iid}{\sim} N(0, \sigma^2), \\ \eta &:= \sum_{i=1}^{l} d_i \phi_i + \eta_1. \end{aligned} \tag{2.61}$$

Here it is assumed that η_1 is a Gaussian process with mean 0 and, for some $b > 0$, a covariance structure of the form

$$\mathbb{E}\left[\eta_1(\mathbf{x}_1), \eta_1(\mathbf{x}_2)\right] = bR_J(\mathbf{x}_1, \mathbf{x}_2),$$

and R_J the reproducing kernel on a *RKHS* . *For example, R_J may be expression* (2.30) *from the thin plate splines setting (Section 2.1.1) or R_J may use* (2.30) *as in the tensor thin plate spline setting (Section* 2.1.2*).* Some priors on the parameters $\{d_i\}_{i=1}^l$ are needed to be specified and description of the functions $\{\phi_i\}_{i=1}^l$ require further explanation; we explain all details in Section 3. The form of η in this model is justified by Theorem 5 and Propositions 6 and 7 with $k = n$, which together state that the mean of the full conditional posterior of $\eta(\mathbf{x})$ under this model is the unique solution to (1.1). We provide more details on this Bayes model in Chapter 3.

The parameters $\{d_i\}_{i=1}^{l}$ are essential for the mean function η but they are not needed for the proposal of good smoothing parameters. The nuisance $\{d_i\}_{i=1}^{l}$ can be eliminated if the likelihood of the transformation $\mathfrak{y} = F_2^\intercal \mathbf{y}$ is used, provided that S is full column rank. Here we will be using the same notation as in Section 2, with F_2 is given by (2.8). Observe that

$$\begin{aligned}
\mathfrak{y} &:= F_2^\intercal \mathbf{y} \\
&= F_2^\intercal S \begin{pmatrix} d_1 \\ \vdots \\ d_l \end{pmatrix} + F_2^\intercal \begin{pmatrix} \eta(x_1) \\ \vdots \\ \eta(x_n) \end{pmatrix} + F_2^\intercal \epsilon \\
&= F_2^\intercal F_1 \tilde{R} \begin{pmatrix} d_1 \\ \vdots \\ d_l \end{pmatrix} + F_2^\intercal \begin{pmatrix} \eta(x_1) \\ \vdots \\ \eta(x_n) \end{pmatrix} + F_2^\intercal \epsilon \qquad (2.62) \\
&= F_2^\intercal \begin{pmatrix} \eta(x_1) \\ \vdots \\ \eta(x_n) \end{pmatrix} + F_2^\intercal \epsilon, \qquad (2.63)
\end{aligned}$$

where equality (2.62) is by the QR-decomposition of S (equation (2.8)), while equality (2.63) is because $F_2^\intercal F_1 = 0$. Hence

$$\mathfrak{y} \sim N(\mathbf{0}, bF_2^\intercal Q F_2 + \sigma^2 I_{n-l}),$$

where equation (2.10) was used for the expression of the variance. Then, the minus log likelihood of $\mathfrak{y}$ is, up to a constant, given by

$$\begin{aligned}
&\frac{1}{2}\mathfrak{y}^\intercal \left(bF_2^\intercal Q F_2 + \sigma^2 I_{n-l}\right)^{-1} \mathfrak{y} + \frac{1}{2}\log\left|bF_2^\intercal Q F_2 + \sigma^2 I_{n-l}\right| \\
&= \frac{1}{2}\mathfrak{y}^\intercal \left(bQ^\star + \sigma^2 I_{n-l}\right)^{-1} \mathfrak{y} + \frac{1}{2}\log\left|bQ^\star + \sigma^2 I_{n-l}\right| \\
&= \frac{1}{2b}\mathfrak{y}^\intercal \left(Q^\star + \frac{\sigma^2}{b} I_{n-l}\right)^{-1} \mathfrak{y} + \frac{1}{2}\log\left\{ b^{n-l} \left| Q^\star + \frac{\sigma^2}{b} I_{n-l} \right| \right\} \\
&= \frac{1}{2b}\mathfrak{y}^\intercal \left(Q^\star + n\lambda I_{n-l}\right)^{-1} \mathfrak{y} + \frac{1}{2}\log\left\{ \left| Q^\star + n\lambda I_{n-l} \right| \right\} + \frac{n-l}{2}\log b \qquad (2.64)
\end{aligned}$$

where $Q^\star = F_2^\intercal Q F_2$ and $n\lambda = \frac{\sigma^2}{b}$. Expression (2.64) can be easily minimized with respect to b by taking the derivatives, equating to 0 and solving. The minimizer of (2.64) is obtained to be

$$\hat{b}(\lambda) = \frac{\mathfrak{y}^\intercal \left(Q^\star + n\lambda I_{n-l}\right)^{-1} \mathfrak{y}}{n-l}. \tag{2.65}$$

Further algebra leads to

$$\begin{aligned}
\hat{b}(\lambda) &= \frac{\mathbf{y}^\intercal \left\{ n\lambda F_2 \left(F_2^\intercal Q F_2 + n\lambda I_{n-l}\right)^{-1} F_2^\intercal \right\} \mathbf{y} (n\lambda)^{-1}}{n-l} \\
&= \frac{\mathbf{y}^\intercal \left[I_n - \left\{ I_n - n\lambda F_2 \left(F_2^\intercal Q F_2 + n\lambda I_{n-l}\right)^{-1} F_2^\intercal \right\} \right] \mathbf{y}}{n\lambda(n-l)} \\
&= \frac{\mathbf{y}^\intercal \left(I_n - A(\lambda)\right) \mathbf{y}}{n\lambda(n-l)}, \qquad (2.66)
\end{aligned}$$

where equation (2.21) was used to obtain (2.66).

We can now see the reason for choosing the form of the hyper-parameter b from the Bayesian model (2.61) as $n\lambda = \frac{\sigma^2}{b}$; only with this form do we obtain an interpretation of $\lambda = \frac{\sigma^2}{nb}$ as the smoothing parameter from the minimization problem of the functional (2.1) which is induced because the smoothing matrix (2.21) associated with (1.1) is used.

Hence, the smoothing parameter λ may be proposed using the minimizer of the profile log likelihood of λ in model (2.61):

$$\begin{aligned}
&\frac{1}{2\hat{b}(\lambda)} \mathfrak{y}^\intercal \left(Q^\star + n\lambda I_{n-l}\right)^{-1} \mathfrak{y} + \frac{1}{2} \log\left\{ \left| Q^\star + n\lambda I_{n-l} \right| \right\} + \frac{n-l}{2} \log \hat{b}(\lambda) \\
&\quad = \frac{1}{2} \left((n-l) + (n-l) \log \left\{ \left| Q^\star + n\lambda I_{n-l} \right|^{1/(n-l)} \hat{b}(\lambda) \right\} \right) \qquad (2.67) \\
&\quad = \frac{n-l}{2} \left(1 + \log \left\{ \left| Q^\star + n\lambda I_{n-l} \right|^{1/(n-l)} \hat{b}(\lambda) \right\} \right). \qquad (2.68)
\end{aligned}$$

Note equation (2.65) was used for (2.67). It is straightforward to see from (2.68) that the minimizer of the profile log likelihood of λ is the same as the minimizer of the function $\mathcal{M}$ defined as

$$\mathcal{M}(\lambda) := \frac{(n-l)\hat{b}(\lambda)}{\left| Q^\star + n\lambda I_{n-l} \right|^{-1/(n-l)}}.$$

Further algebra on $\mathcal{M}$ leads to

$$\mathcal{M}(\lambda) = \frac{1}{n\lambda} \frac{\mathbf{y}^\intercal \left(I_n - A(\lambda)\right) \mathbf{y}}{\left| \left(Q^\star + n\lambda I_{n-l}\right)^{-1} \right|^{1/(n-l)}} \tag{2.69}$$

$$= \frac{1}{n\lambda} \frac{\mathbf{y}^\intercal (I_{n-l} - A(\lambda))\, \mathbf{y}}{\left| F_2 \left(Q^\star + n\lambda I_{n-l}\right)^{-1} F_2^\intercal \right|_+^{1/(n-l)}} \tag{2.70}$$

$$= \frac{1}{n\lambda} \frac{\mathbf{y}^\intercal (I_{n-l} - A(\lambda))\, \mathbf{y}}{(n\lambda)^{-1} \left| n\lambda F_2 \left(Q^\star + n\lambda I_{n-l}\right)^{-1} F_2^\intercal \right|_+^{1/(n-l)}}$$

$$= \frac{\mathbf{y}^\intercal (I - A(\lambda))\, \mathbf{y}}{|I - A(\lambda)|_+^{1/(n-l)}}, \tag{2.71}$$

where $|B|_+$ denotes the product of the positive eigenvalues of B. (2.69) above was obtained using (2.66), the equality needed for the denominator in (2.70) was proven in Wahba (1985), while (2.71) is obtained with the help of (2.21). We have obtained a handy expression of $\mathcal{M}$ as:

$$\mathcal{M}(\lambda) \propto \frac{\mathbf{y}^\intercal (I - A(\lambda))\, \mathbf{y}}{|I - A(\lambda)|_+^{1/(n-l)}}. \tag{2.72}$$

The minimizer λ_m of (2.72) is called by Wahba (1985) the *Restricted Maximum Likelihood* estimate of λ, we will denote this as RML.

The corresponding variance estimate using $n\lambda_m = \sigma_m^2 / b(\lambda_m)$ is

$$\sigma_m^2 = \frac{\mathbf{y}^\intercal (I - A(\lambda_m))\, \mathbf{y}}{n - l}. \tag{2.73}$$

2.3.4 Smoothing parameter as the ratio of variances

A simple solution to the problem of choosing the smoothing parameter λ when it is not needed to choose the bandwidth parameters $\{\theta\}_{i=1}^p$, can be obtained when considering the linear mixed model in Section 2.2.2. The expression for $n\lambda = \frac{\sigma_c^2}{\sigma_e^2}$ is observed when comparing (2.52) and (2.50). In other words, $n\lambda$ is the ratio between the variance of the errors at the level of the response and the variance of the random effects of the linear mixed model in Section 2.2.2. Given estimators of the two variances, the ratio of these would provide an estimator of $n\lambda$. In a Bayesian model approach, one could assign priors to σ_c^2 and σ_e^2.

CHAPTER 3. BAYESIAN MODEL USING THE APPROXIMATED SOLUTION FOR THE PENALIZED LEAST SQUARES MINIMIZATION PROBLEM

In this chapter we describe a Bayesian regression model with the property that the mean of the full conditional distribution of the estimated function is the approximated solution to the penalized least squared minimization problem (1.1); details about the approximated solution are given in Section 2.2. A Bayesian method with this property was proposed first in Wahba (1978) for the simple regression problem. In Wahba (1983), the authors discuss the properties of credible sets when the Bayes approach is used in the bivariate regression model. In those early papers, the authors obtained the exact solution to the penalized least square minimization problem; the approximated solution as mean posterior was proposed more recently in Kim and Gu (2004). All the Bayesian models described here, conditional on the smoothing parameters and the covariance of the error-response, have the property that when $\{\mathbf{z}_i\}_{i=1}^k = \{\mathbf{x}_i\}_{i=1}^n$, the resulting model is the same model as described by Wahba (1983) and its full conditional posterior mean provides an exact solution to (1.1).

We present a Bayesian model with the property that the mean of the full conditional posterior distribution is the same regression estimators as in Kim and Gu (2004) and require the same computational effort. The method we propose has an important practical advantage over the Kim *et.al.* model in the error in variable regression case (Section 4). Kim *et.al.* treat the regression function as a Gaussian process while we formulate it as linear combination of a set of known fixed basis functions. This means that extending Kim's method to the measurement error case requires that in each iteration, we save the sampling points $\{\mathbf{x}_i\}_{i=1}^n$ at which the regression function is estimated, as well as the estimates of the regression function itself. This is impractical because we have to choose from the beginning all the points $\{\chi_i\}_{i=1}^N$ where we

want to estimate the regression function and save the grid and the estimated function from every MCMC iteration. To implement the approach we propose, on the other hand, we only need to save the posterior MCMC samples of the coefficients $\begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix}$ and use these with the fixed basis evaluated at any new points $\{\chi_i\}_{i=1}^N$ (or sampling points $\{\mathbf{x}_i\}_{i=1}^n$). In Section 4 we use this methodology to address the measurement error problem since reducing as much as possible the computational burden is important.

First we introduce the approach proposed by Kim *et.al.* to show its equivalence to the basis model approach that we propose in Section 3.1. In Section 3.2 we describe the models considered in our fitting methodology which are compared in the simulation study in Section 3.3.

3.1 First Theoretical Results

We assume that we have a training set $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{X} \times \mathbb{R}$ (for example $\mathbb{X} = \mathbb{R}^d$), and we let $\{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n$ denote a set of knots that may be chosen as described in Section 2.2.3. Consider the model

$$\begin{aligned} y_i &= \eta(\mathbf{x}_i) + \epsilon_i, \\ \epsilon_i &\overset{iid}{\sim} N(0, \sigma^2). \end{aligned} \tag{3.1}$$

Kim *et.al.* methodology uses the decomposition (2.48) of $\eta \in \mathcal{H}^\star$ to estimate η as the solution to (1.1) in a Bayesian context; they write $\eta = \eta_0 + \eta_1$ where $\eta_0 \in \mathcal{N}_J$ and $\eta_1 \in \mathcal{H}^\star \ominus \mathcal{N}_J$. Theorem 5 summarizes some of the results Kim and Gu (2004) of interest. Theorem 5 uses some of the ideas from the preliminaries in Section 2.

Theorem 5 (Posterior Process, Kim and Gu (2004))
In the context of Section 2.2 and model (3.1), let $\{\psi_i\}_{i=1}^l$ be a basis of $\mathcal{N}_J$. Let $\lambda > 0$, and let $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$, $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ with $S_{i,j} = \psi_j(\mathbf{x}_i)$ full column rank, $R \in \mathcal{M}_{n\times k}(\mathbb{R})$, $R_{i,j} = R_J(\mathbf{x}_i, \mathbf{z}_j)$, $M = RQ^+R^\intercal + n\lambda I_n$ and define $b = \frac{\sigma^2}{n\lambda}$. Consider the unique decomposition $\eta = \eta_0 + \eta_1 \in \mathcal{H}^\star$. Let η_0 have a improper prior in $\mathcal{N}_J$ and let η_1 follow a mean zero Gaussian process prior in

$\mathcal{H}^\star \ominus \mathcal{N}_J$ with covariance function

$$\mathbb{E}(\eta_1(\mathbf{x}), \eta_1(\mathbf{y})) = b \begin{pmatrix} R_J(\mathbf{x}, \mathbf{z}_1) \\ \vdots \\ R_J(\mathbf{x}, \mathbf{z}_k) \end{pmatrix}^\intercal Q^+ \begin{pmatrix} R_J(\mathbf{z}_1, \mathbf{y}) \\ \vdots \\ R_J(\mathbf{z}_k, \mathbf{y}) \end{pmatrix},$$

were Q^+ is the Moore-Penrose inverse (Definition 32) of

$$Q = \begin{pmatrix} R_J(\mathbf{z}_1, \mathbf{z}_1) & R_J(\mathbf{z}_1, \mathbf{z}_2) & \cdots & R_J(\mathbf{z}_1, \mathbf{z}_k) \\ R_J(\mathbf{z}_2, \mathbf{z}_1) & R_J(\mathbf{z}_2, \mathbf{z}_2) & \cdots & R_J(\mathbf{z}_2, \mathbf{z}_k) \\ \vdots & \vdots & \ddots & \vdots \\ R_J(\mathbf{z}_k, \mathbf{z}_1) & R_J(\mathbf{z}_k, \mathbf{z}_2) & \cdots & R_J(\mathbf{z}_k, \mathbf{z}_k) \end{pmatrix}. \tag{3.2}$$

Then, the posterior distribution $[\eta|\mathbf{y}, \sigma^2]$ is a Gaussian process with

$$\begin{aligned} \mathbb{E}\left[\eta(\mathbf{x})|\,\mathbf{y}, \sigma^2\right] &= \sum_{i=1}^{l} \hat{d}_i \psi_i(\mathbf{x}) + \sum_{i=1}^{k} \hat{c}_i R_J(\mathbf{z}_i, \mathbf{x}) \\ &= \psi^\intercal \hat{\mathbf{d}} + \xi^\intercal \hat{\mathbf{c}}, \\ b^{-1} Var[\eta(\mathbf{x})|\mathbf{y}] &= \xi^\intercal Q^+ \xi + \psi^\intercal \left(S^\intercal M^{-1} S\right)^{-1} \psi - 2\psi^\intercal \tilde{\mathbf{d}} - \xi^\intercal \tilde{\mathbf{c}}, \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} \hat{\mathbf{c}} &= Q^+ R^\intercal \left(M^{-1} - M^{-1} S \left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1}\right) \mathbf{y}, \\ \hat{\mathbf{d}} &= \left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1} \mathbf{y}, \\ \tilde{\mathbf{c}} &= Q^+ R^\intercal \left(M^{-1} - M^{-1} S \left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1}\right) R Q^+ \xi, \\ \tilde{\mathbf{d}} &= \left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1} R Q^+ \xi. \end{aligned} \tag{3.4}$$

Furthermore $\hat{\mathbf{d}}$ and $\hat{\mathbf{c}}$ satisfy equations (2.51) and thus the mean posterior is the approximated solution to the penalized least square minimization problem (1.1) described in Section 2.2.

Expressions (3.3), for known σ^2, uniquely describe the posterior distribution $[\eta(\mathbf{x}_i)|\mathbf{y}, \lambda, \mathbf{X}, \sigma^2]$ for any sampling point $\mathbf{x}_i$, or they can be used in the posterior predictive distribution of $\eta(\chi)$ for any new $\chi \in \mathbb{X}$. Kim *et.al* provide an external estimator $\hat{\sigma}^2$ of σ^2 using the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, and then continue as if the variance $\hat{\sigma^2} = \sigma^2$ is known, and provide point estimates of $\eta(\chi)$ using the normal distribution with parameters given by (3.3) and (3.4). This approach may be classified as empirical Bayes.

Now lets assume model (3.1) with η having the form

$$\eta(\mathbf{x}) = \sum_{i=1}^{l} d_i \psi_\nu(\mathbf{x}) + \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \mathbf{x})$$

as in the setting of Section 2.2. In the notation of Theorem 5 we have $\eta_0 = \sum_{i=1}^{l} d_i \psi_i$ and $\eta_1 = \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \cdot)$. From the discussion in Chapter 2, the form of η is justified as the form of the approximated solution to the penalized least square minimization problem (1.1) for a chosen penalty term (2.2) (and known $\{\theta_i\}_{i=1}^{p}$) and a fixed known value λ. The basis functions $\{\psi_i\}_{i=1}^{l}$ of the space $\mathcal{N}_J$ and the functions $\{R_J(\mathbf{z}_i, \cdot)\}_{i=1}^{k}$ are known and fixed and thus the unknown parameters are $\{d_i\}_{i=1}^{l}$ and $\{c_i\}_{i=1}^{k}$; one may assign priors to these parameters as described in the next Proposition.

Proposition 6 (A First Bayesian Model)

In the context of Section 2.2, let $\{\psi_i\}_{i=1}^{l}$ be a basis of $\mathcal{N}_J$. Consider $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \subset \mathbb{X} \times \mathbb{R}$ a training set, let $\mathbf{Z} := \{\mathbf{z}_i\}_{i=1}^{k} \subset \{\mathbf{x}_i\}_{i=1}^{n} =: \mathbf{X}$, $\mathbf{d} := (d_1\, d_2\, \cdots\, d_l)^\intercal$, $\mathbf{c} := (c_1\, c_2\, \cdots\, c_k)^\intercal$. Consider the model

$$\begin{aligned} y_i &= \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i) + \epsilon_i, \\ \eta_{\binom{\mathbf{d}}{\mathbf{c}}} &= \sum_{i=1}^{l} d_i \psi_i + \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \cdot) \\ \epsilon_i &\overset{iid}{\sim} N_1(0, \sigma^2). \end{aligned}$$

Let $\lambda > 0$, and $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ with entries $Q_{i,j} = R_J(\mathbf{z}_i, \mathbf{z}_j)$, $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ with $S_{i,j} = \psi_j(\mathbf{x}_i)$ full column rank, $R \in \mathcal{M}_{n\times k}(\mathbb{R})$, $R_{i,j} = R_J(\mathbf{x}_i, \mathbf{z}_j)$, $M = RQ^{+}R^\intercal + n\lambda I_n$ and define $b = \frac{\sigma^2}{n\lambda}$. Consider the priors

$$\begin{aligned} d_i &\overset{iid}{\sim} 1, \\ \mathbf{c}|\sigma^2 &\sim N_l\left(\mathbf{0}, b\, Q^{+}\right), \\ \sigma^2 &\sim Inv - Gamma(A_\epsilon, B_\epsilon), \\ \mathbf{d} &\perp \mathbf{c}, \\ \mathbf{d} &\perp \sigma^2 \\ \binom{\mathbf{d}}{\mathbf{c}} &\perp (\epsilon_1 \cdots \epsilon_n)^\intercal, \end{aligned}$$

$$\sigma^2 \perp \epsilon_i,\ i \in \{1, \cdots, n\}.$$

Then the posterior of the parameters exists and the full conditional posteriors are

- $\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)|\mathbf{y},\ \sigma^2,\ b,\ \mathbf{X} \sim N_{l+k}\left(\mu_{\mathbf{dc}}, b\mathbf{\Sigma_{dc}}\right)$, *where*

$$\mu_{\mathbf{dc}} = \begin{pmatrix} (S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1} \\ Q^+ R^\intercal M^{-1}(I - S(S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1}) \end{pmatrix} \mathbf{y} \tag{3.5}$$

$$\mathbf{\Sigma_{dc}} = \begin{pmatrix} (S^\intercal M^{-1} S)^{-1} & -(S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1} R Q^+ \\ -Q^+ R^\intercal M^{-1} S (S^\intercal M^{-1} S)^{-1} & Q^+ - Q^+ R^\intercal \{M^{-1} - M^{-1} S (S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1}\} R Q^+ \end{pmatrix} \tag{3.6}$$

- $\sigma^2|\mathbf{y},\ \left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right),\ \mathbf{X} \sim Inv - Gamma\left(A_\epsilon + \frac{1}{2}n,\ \left[B_\epsilon^{-1} + \frac{1}{2}\sum_{i=1}^n (y_i - \eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}(\mathbf{x}_i))^2\right]^{-1}\right)$.

Proof.

Denote $\mathbf{\Theta} = \left(\mathbf{d}^\intercal\ \mathbf{c}^\intercal\ \sigma^2\right)^\intercal$. The posterior density of $\mathbf{\Theta}$ can be expressed as

$$[\mathbf{\Theta}|\mathbf{y},\ \mathbf{X}] \propto \left[\mathbf{y}|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right),\ \sigma^2,\ \mathbf{X}\right] \times \left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\sigma^2,\ \mathbf{X}\right] \times [\sigma^2],$$

thus the posterior is proper if the right hand side of the previous expression is proper. The distribution $\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\mathbf{y}, \sigma^2,\ \mathbf{X}\right] \propto \left[\mathbf{y}|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right),\ \sigma^2,\ \mathbf{X}\right] \times \left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\sigma^2,\ \mathbf{X}\right]$ may not integrate to 1, but the rest of the distributions are proper. This distribution can be shown to be proper by considering first the model with proper prior $d_i \overset{iid}{\sim} N(0, \tau^2)$ with the rest of the priors kept the same. For $\tau^2 \to \infty$ we can then prove that $\left[\left(\begin{smallmatrix}\mathbf{d}_\tau\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\mathbf{y}, \sigma^2\,\mathbf{X}\right]$ converges in distribution. This is shown in Proposition 74.

To compute the full conditional posterior $\left[\sigma^2\middle|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right), \mathbf{X}, \mathbf{y}\right]$ we recall that

$$[\mathbf{\Theta}|\mathbf{y},\ \mathbf{X}] \propto \left[\,\mathbf{y}\,\middle|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right), \sigma^2,\ \mathbf{X}\right] \left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\sigma^2,\ b,\ \mathbf{X}\right] \left[\sigma^2|\mathbf{X}\right],$$

and obtain

$$\left[\sigma^2\middle|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right), \mathbf{X}, \mathbf{y}\right] \propto \left[\,\mathbf{y}\,\middle|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right), \sigma^2,\ \mathbf{X}\right] \left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\sigma^2,\ b,\ \mathbf{X}\right] \left[\sigma^2|\mathbf{X}\right]. \tag{3.7}$$

But $\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\sigma^2,\ b,\ \mathbf{X}\right]$ depends on σ^2 only through the expression σ^2/b and

$$\frac{\sigma^2}{b} = \frac{\sigma^2}{\sigma^2/n\lambda} = n\lambda.$$

Therefore, we have $\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\sigma^2,\ b,\ \mathbf{X}\right] = \left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\,\lambda,\ \mathbf{X}\right]$ which indicates that the conditional distribution is independent of σ^2. By (3.7), we have

$$\left[\sigma^2\middle|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right), \mathbf{X}, \mathbf{y}\right] \propto \left[\,\mathbf{y}\,\middle|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right), \sigma^2,\ \mathbf{X}\right] \left[\sigma^2|\mathbf{X}\right]$$

$$\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}(\mathbf{x}_i)\right)^2-\frac{1}{2B_\epsilon\sigma^2}\right)\times\sigma^{-2(n/2+A_\epsilon+1)}$$

$$\propto Inv-Gamma\left(A_\epsilon+\frac{1}{2}n,\left[B_\epsilon^{-1}+\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}(\mathbf{x}_i)\right)^2\right]^{-1}\right).$$

■

Proposition 6 under-girds a main theoretical result proposed here and we will use it for the rest of this paper. Kim's *et.al.* results, Theorem 5, is related to our own via the point estimates of η in any $\chi \in \mathbb{R}^d$. We claim that the full conditional posterior distribution of η in Proposition 6, as a process, is the same as the one proposed by *Kim, et.al.* from Theorem 5. We state and prove this result in Proposition 7.

Proposition 7 (Equivalence Bayesian Models)

In the context of Proposition 6, $\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}$ *is a Gaussian process with mean and covariance functions*

$$\begin{aligned}
\mathbb{E}\left[\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x})\right] =& \boldsymbol{\Psi}^\intercal(\mathbf{x})\hat{\mathbf{d}}+\boldsymbol{\Xi}^\intercal(\mathbf{x})\hat{\mathbf{c}},\\
b^{-1}Cov\left[\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x}),\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{y})\right] =& \boldsymbol{\Xi}(\mathbf{x})^\intercal Q^+\boldsymbol{\Xi}(\mathbf{y})+\boldsymbol{\Psi}(\mathbf{x})^\intercal\left(S^\intercal M^{-1}S^\intercal\right)^{-1}\boldsymbol{\Psi}(\mathbf{x})\\
&-\left[\boldsymbol{\Psi}(\mathbf{x})^\intercal\tilde{\mathbf{d}}(\mathbf{y})+\boldsymbol{\Psi}(\mathbf{y})^\intercal\tilde{\mathbf{d}}(\mathbf{x})\right]-\boldsymbol{\Xi}(\mathbf{x})^\intercal\tilde{\mathbf{c}}(\mathbf{y})\\
b^{-1}Var\left[\eta_{\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)}\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x})\right] =& \boldsymbol{\Xi}(\mathbf{x})^\intercal Q^+\boldsymbol{\Xi}(\mathbf{x})+\boldsymbol{\Psi}^\intercal(\mathbf{x})\left(S^\intercal M^{-1}S\right)^{-1}\boldsymbol{\Psi}(\mathbf{x})\\
&-2\boldsymbol{\Psi}(\mathbf{x})^\intercal\tilde{\mathbf{d}}(\mathbf{x})-\boldsymbol{\Xi}(\mathbf{x})^\intercal\tilde{\mathbf{c}}(\mathbf{x}),
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\Psi}(\mathbf{x}) &= (\psi_1(\mathbf{x})\ \cdots\ \psi_l(\mathbf{x}))^\intercal\\
\boldsymbol{\Xi}(\mathbf{x}) &= (R_J(\mathbf{z}_1,\mathbf{x})\ \cdots\ R_J(\mathbf{z}_k,\mathbf{x}))^\intercal\\
\hat{\mathbf{c}} &= Q^+R^\intercal\left(M^{-1}-M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\mathbf{y}, \qquad (3.8)\\
\hat{\mathbf{d}} &= \left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y}, \qquad (3.9)\\
\tilde{\mathbf{c}}(\mathbf{x}) &= Q^+R^\intercal\left(M^{-1}-M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)RQ^+\boldsymbol{\Xi}(\mathbf{x}),\\
\tilde{\mathbf{d}}(\mathbf{x}) &= \left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}RQ^+\boldsymbol{\Xi}(\mathbf{x}).
\end{aligned}$$

The Gaussian process $\eta\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}$ *is exactly the same process as in Theorem 5 described by (3.3) and (3.4).*

Proof. By Proposition 75 with a slightly different notation. ■

We have shown that the full conditional posterior distribution of the regression function η evaluated at any $\chi \in \mathbb{R}^d$, conditional on σ and the smoothing parameters λ, $\{\theta_i\}_{i=1}^p$, is the same as the full conditional posterior of the Gaussian process η from Theorem 5 evaluated at $\chi \in \mathbb{R}^d$. In practical terms, this means that point estimators and credible intervals for $\eta(\chi)$ are the same for both models conditional on σ^2, λ, $\{\theta_i\}_{i=1}^p$ and $\{\mathbf{x}_i\}_{i=1}^n$. The advantage of our results over Kim, *et.al.* is in terms of storage, prediction, and the extension to the measurement error case.

In order to predict using the model from Proposition 6 we only need to save the values $\hat{\mathbf{d}}$ and $\hat{\mathbf{c}}$ from an MCMC sampler and then use them to evaluate the basis of functions $\{\psi_i\}_{i=1}^l$ and $\{R_J(\mathbf{z}_i,\cdot)\}_{i=1}^k$ to simulate from the marginal posterior predictive distribution $[\eta(\chi)|\mathbf{y},\mathbf{X},\lambda]$ for any new χ. On the other side, to simulate from the posterior predictive of $\eta(\chi)$ using the model in Theorem 5 we have to draw from the posterior distribution as we estimate the model parameters. This limits the points where η can be estimated to the sets of points $\{\chi_i\}_{i=1}^N$ defined before the fitting of the model.

For the sake of completing the arguments we now show that $\hat{\mathbf{d}}$ and $\hat{\mathbf{c}}$ from (3.4), $\mu_{\mathbf{dc}}$ in (3.5) and $\hat{\mathbf{d}}$, $\hat{\mathbf{c}}$ from (3.8) and (3.9), which are the same algebraic expressions, satisfy equations (2.51). Therefore the mean of $\eta(\chi)$ of the Gaussian process in Theorem 5 and the mean of the full conditional posterior of η from Proposition 6 as functions of $\mathbf{x} \in \mathbb{R}^d$ solve the regularized minimization problem (1.1) in the space $\mathcal{H}^\star$.

Proposition 8 (Full Conditional Posterior Mean as Smoothing Splines) *In the setting of Theorem 5 or Proposition 6, let S be full column rank $n \times l$ matrix. The mean of the full conditional posterior of $\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)$ described by (3.5), and the vector $\left(\begin{smallmatrix}\hat{\mathbf{d}}\\ \hat{\mathbf{c}}\end{smallmatrix}\right)$ from (3.8) – (3.9) satisfy equations (2.51) and thus the Bayesian point estimator of the Gaussian process η in Theorem 5 an the full conditional mean in Proposition 6 are the unique solution to (1.1) in $\mathcal{H}^\star$.*

Proof.

The first expression of (3.5) satisfies equations (2.51) because:

$$\begin{aligned}
S^\intercal S\hat{\mathbf{d}}+S^\intercal R\hat{\mathbf{c}} - S^\intercal\mathbf{y} &= S^\intercal S\left[\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y}\right] \\
&\quad + S^\intercal R\left[Q^+R^\intercal\left(M^{-1}+M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\mathbf{y}\right] - S^\intercal\mathbf{y} \\
&= S^\intercal\left(RQ^+R^\intercal\right)M^{-1}\mathbf{y} + S^\intercal\left(I_n - RQ^+R^\intercal M^{-1}\right)S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y} - S^\intercal\mathbf{y} \\
&= S^\intercal\left(RQ^+R^\intercal M^{-1}-I_n\right)\mathbf{y} - S^\intercal\left(RQ^+R^\intercal M^{-1}-I_n\right)S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y} \\
&= S^\intercal\left(RQ^+R^\intercal M^{-1}-I_n\right)\left(I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\mathbf{y} \\
&= S^\intercal\left(-n\lambda M^{-1}\right)\left(I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\mathbf{y} \\
&= -\ n\lambda\left[S^\intercal M^{-1} - \left(S^\intercal M^{-1}S\right)\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= -\,n\lambda\left[S^\intercal M^{-1} - S^\intercal M^{-1}\right]\mathbf{y} \\
&= \mathbf{0}.
\end{aligned}$$

That the second expression of (3.5) satisfies equations (2.51) follows as:

$$\begin{aligned}
R^\intercal S\hat{\mathbf{d}}+\left(R^\intercal R+n\lambda Q\right)\hat{\mathbf{c}} - R^\intercal\mathbf{y} &= R^\intercal S\left[\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y}\right] - R^\intercal\mathbf{y} \\
&\quad + \left(R^\intercal R+n\lambda Q\right)\left[Q^+R^\intercal M^{-1}(I - S(S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1})\right] \\
&= R^\intercal S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y} - R^\intercal\mathbf{y} \\
&\quad + R^\intercal RQ^+R^\intercal M^{-1}\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&\quad + n\lambda QQ^+R^\intercal M^{-1}\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= R^\intercal S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\mathbf{y} - R^\intercal\mathbf{y} \\
&\quad + R^\intercal RQ^+R^\intercal M^{-1}\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&\quad + n\lambda R^\intercal M^{-1}\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \qquad (3.10) \\
&= \left[R^\intercal\left(RQ^+R^\intercal M^{-1}\right) + n\lambda R^\intercal M^{-1} - R^\intercal\right]\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= \left[R^\intercal\left(I_n - n\lambda M^{-1}\right) + n\lambda R^\intercal M^{-1} - R^\intercal\right]\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= \left[R^\intercal - n\lambda R^\intercal M^{-1} + n\lambda R^\intercal M^{-1} - R^\intercal\right]\left[I_n - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
&= \mathbf{0}.
\end{aligned}$$

Equality (3.10) is obtained using $QQ^{+}R^{\intercal} = R^{\intercal}$ which can be proven as follow. Let

$$\xi(\mathbf{x}) = (R_J(\mathbf{z}_1, \mathbf{x}) \cdots R_J(\mathbf{z}_k, \mathbf{x}))^{\intercal}.$$

Notice that by definition of the generalized inverse Q^{+}, we have $QQ^{+}Q = Q$, therefore QQ^{+} is the projection matrix on the column space of Q; if we prove that $\xi(\mathbf{x})$ is in the column space of Q we would have proven $QQ^{+}\xi(\mathbf{x}) = \xi(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^l$, in particular we would have proven that $QQ^{+}R^{\intercal} = R^{\intercal}$. $\xi(\mathbf{x})$ is in the column space of Q by Proposition 46, and thus we have (3.10).

We have shown that $\hat{\mathbf{d}}$ and $\hat{\mathbf{c}}$ satisfy equation (2.51), but this by itself only shows that they are critical points of (2.50). As described in Section 2.2.1, there may be multiple critical points to (2.50) but all the critical points would produce a solution to (1.1) through expressions (2.49), along with the expected value in (3.3) of $\mathbb{E}(\eta(\mathbf{x}))$, or $\mathbb{E}\left[\eta\binom{\mathbf{d}}{\mathbf{c}}\Big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x})\right]$ from Proposition 6. Such solution would be in the space of functions $\{\eta = \sum_{i=1}^{l} d_i\psi_i + \sum_{i=1}^{k} R_J(\mathbf{z}_i, \cdot)\}$. That the solution is unique in $\mathcal{H}^{\star}$ is concluded by the representer Theorem (Theorem 52) which requires that S is full column rank. ■

Another interpretation of the deterministic function defined by the mean of the Gaussian process from the posterior predictive was already mentioned but it may not be evident in this setting. It was argued in Section 2.2.1 using Theorem 51 that the minimizer function, in this case $\mathbb{E}\left[\eta\binom{\mathbf{d}}{\mathbf{c}}\Big|_{\mathbf{y},\sigma^2,b,\mathbf{X}}(\mathbf{x})\right]$ as function of $\mathbf{x} \in \mathbb{R}^d$, is the best interpolation as measured by the quadratic loss function $\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2$ subject to the constrain $J(f) \leq \rho(\lambda)$ for $\rho(\lambda) > 0$.

3.2 Models

In this Section we describe the Bayesian models that we discuss in this section. In every case we assume that we have a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ is available, and a set of knots $\{(\mathbf{z}_i)\}_{i=1}^{k}$, $k << n$ with similar distribution as $\{(\mathbf{x}_i)\}_{i=1}^{n}$ (see Section 2.2.3). In the decomposition of the space $\mathcal{H}^{\star}$ (2.48), the Hilbert subspace $\mathcal{N}_J$ is of finite dimension with basis $\{\psi_i\}_{i=1}^{l}$ and orthonormal basis $\{\phi_i\}_{i=1}^{l}$. We have available a semi kernel $R_J : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ in the space $\mathcal{H}^{*}$.

3.2.1 Bayesian regression model using thin plate splines

The thin plate spline regression setting, described in Section 2.1.1, will be used here. Of particular importance is the evaluation of the reproducing kernel function $R_{J_m^d}$ (for simplicity R_J). Proposition 6 describes the most troublesome part of the model in this section, namely the priors and conditional posterior of $\mathbf{c}$ and $\mathbf{d}$ given the variance at the observation level σ^2 and the bandwidth parameter $\lambda > 0$ are required to solve (1.1).

For $\eta \in \mathcal{H}^*$ consider the model

$$y_i = \eta(\mathbf{x}_i) + \epsilon_i$$
$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Here $\eta = \sum_{j=1}^{l} d_j \phi_j + \sum_{j=1}^{q} c_j R_J(\mathbf{z}_j, \cdot)$ by hypothesis. Using the same notation as in Proposition 6, consider the priors on the parameters

$$\mathbf{d} \sim 1,$$

$$\mathbf{c}|\sigma^2, \lambda \sim N_k\left(\mathbf{0}, \frac{\sigma^2}{n\lambda} Q^+\right),$$

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X}, \sigma^2) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{U}(x|\mathbf{y}, \mathbf{X}, \sigma^2)\right\} dF_{\mathbf{y}|\mathbf{X}}(\mathbf{y}), \tag{3.11}$$

$$\sigma^2 \sim Inv - Gamma(A_\epsilon, B_\epsilon),$$

$$\mathbf{d} \perp \mathbf{c}, \quad \mathbf{d} \perp \sigma^2, \quad \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} \perp (\epsilon_1 \cdots \epsilon_n)^\intercal, \quad \sigma^2 \perp \epsilon_i, \; i \in \{1, \cdots, n\},$$

where every distribution is conditional on the observed $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, and A_ϵ and B_ϵ are hyperprior parameters. The prior on λ is conditional on σ^2, but two similar models with λ independent of σ^2 can be proposed. Here the priors on λ would only depend on $\{\mathbf{x}_i\}_{i=1}^n$ and can be chosen as

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X}) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{V}(x|\mathbf{y}, \mathbf{X}, \alpha)\right\} dF_{\mathbf{y}|\mathbf{X}}(\mathbf{y}) \text{ or,} \tag{3.12}$$

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X}) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{M}(x|\mathbf{y}, \mathbf{X})\right\} dF_{\mathbf{y}|\mathbf{X}}(\mathbf{y}); \tag{3.13}$$

for the expression of $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{M}$ see (2.55), (2.59) and (2.72) respectively.

Let $\Theta = \left(\mathbf{d}^\intercal, \mathbf{c}^\intercal, \sigma_\epsilon^2, \lambda\right)$. In principle, we would need to show that the posterior $[\Theta|\mathbf{y}]$ exists because an improper prior was given to $\mathbf{d}$. The formal way to show the existence of the posterior is to propose the proper prior $\mathbf{d} \sim N_l(\mathbf{0}, \tau^2 I_l)$, and follow the proof of Proposition 72 by taking the limit $\tau \to \infty$. The details of the proof are exactly the same as in Proposition 72 with the addition of multiplicative terms independent of τ, where the multiplicative terms correspond to the joint prior distribution of σ^2 and λ using (3.11), (3.12) or (3.13). In this light, the joint posterior distribution of Θ with improper prior on $\mathbf{d}$ exists and is proportional to

$$
\begin{aligned}
[\Theta|\mathbf{y}] &\propto \left[\mathbf{y}|\mathbf{c}, \mathbf{d}, \sigma^2, \lambda\right] \times \left[\mathbf{d}, \mathbf{c}|\lambda, \sigma^2\right] \times \left[\lambda|\sigma^2\right] \times \left[\sigma^2\right] \\
&= \left[\mathbf{y}|\mathbf{c}, \mathbf{d}, \sigma^2\right] \times \left[\mathbf{d}, \mathbf{c}|\lambda, \sigma^2\right] \times \left[\lambda|\sigma^2\right] \times \left[\sigma^2\right] \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \eta(\mathbf{x}_i; \mathbf{d}, \mathbf{c}))^2) - \frac{n\lambda}{\sigma^2}\mathbf{c}^\intercal Q \mathbf{c} - \frac{1}{B_\epsilon \sigma^2}\right\} \times \sigma^{-2(n/2 + A_\epsilon + 1)} \times \left[\lambda|\sigma^2\right].
\end{aligned}
$$

Remark 9

In the expressions above we are using the probability density functions of the respective conditional distributions, or in measure theory terminology, the Radon-Nikodyn derivatives (Athreya and Lahiri (2006)) of the respective probability functions with respect to Lebesgue measures. In this context, $\left[\lambda|\sigma^2\right]$ *would be the Radon-Nikodyn derivative of the measure defined by* (3.11) *with respect to Lebesgue measure. Formally, we would need to prove the existence of such derivative, using for example, the Radon-Nikodym Theorem (Athreya and Lahiri (2006)). If such derivative does not exist then* $[\Theta|\mathbf{y}]$ *can not be analytically expressed as the product of densities, as we did above; instead, the use of the cumulative distributions would be needed. For simplicity in the notation, we keep using the probability density distributions. Furthermore, we do not use the existence of the density* $\left[\lambda|\sigma^2\right]$ *but we will only use that* $\left[\lambda|\sigma^2, \mathbf{y}\right] = \arg\min_{x>0} \mathcal{U}(x|\sigma^2, \mathbf{y})$ *almost surely, by construction of* (3.11)*. Similarly if we use the priors* (3.12) *or* (3.13)*.*

It is now desired to simulate from $[\Theta|\mathbf{y}]$ which can be accomplished using the Gibbs sampler algorithm (Gelman et al., 2014, p. 276 - 278) simulating sequentially from the following full

conditional distributions.

$$\left[(\mathbf{d}^\intercal\ \mathbf{c}^\intercal)^\intercal\,|\lambda,\sigma^2,\mathbf{y}\right] \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\eta(\mathbf{x}_i;\mathbf{d},\mathbf{c}))^2)-\frac{n\lambda}{\sigma^2}\mathbf{c}^\intercal Q\mathbf{c}\right\}$$

$$\sim N_{l+k}\left(\mu_{\mathbf{dc}},\mathbf{\Sigma_{dc}}\right) \text{(following proof Proposition 6), and ,}$$

$$\left[\lambda,\sigma^2|\mathbf{d},\mathbf{c},\mathbf{y}\right] \propto \left[\lambda|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y}\right]\times\left[\sigma^2|\mathbf{d},\mathbf{c},\mathbf{y}\right],$$

for $\mu_{\mathbf{dc}}$ and $\mathbf{\Sigma_{dc}}$ as in (3.5) and (3.6). It is straightforward to simulate from $\left[(\mathbf{d}^\intercal\ \mathbf{c}^\intercal)^\intercal\,|\lambda,\sigma^2,\mathbf{y}\right]$. In order to simulate from $\left[\lambda,\sigma^2|\mathbf{d},\mathbf{c},\mathbf{y}\right]$, first it is needed to simulate from $\left[\sigma^2|\mathbf{d},\mathbf{c},\mathbf{y}\right]$ and using the simulated value σ^2, one can simulate from $\left[\lambda|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y}\right]$. Observe that

$$\left[\sigma^2|\mathbf{d},\mathbf{c},\mathbf{y}\right] \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\eta(\mathbf{x}_i;\mathbf{d},\mathbf{c}))^2-\frac{1}{B_\epsilon\sigma^2}\right\}\times\sigma^{-2(n/2+A_\epsilon+1)}$$

$$\sim Inv-Gamma\left(A_\epsilon+\frac{1}{2}n,\left[B_\epsilon^{-1}+\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\eta(\mathbf{x}_i)\right)^2\right]^{-1}\right)$$

For the distribution $[\lambda|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y}]$ observe that by the law of Total Probability we have

$$\mathbf{P}(\lambda\geq\lambda_0|\sigma^2,\mathbf{y})=\int\mathbf{P}(\lambda\geq\lambda_0|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y})dF_{\mathbf{d},\mathbf{c}}(\mathbf{d},\mathbf{c}),\ \text{then}$$

$$\mathbf{1}\left\{\lambda_0\geq\arg\min_{x>0}\mathcal{U}(x|\mathbf{y},\mathbf{X},\sigma^2)\right\}=\int\mathbf{P}(\lambda\geq\lambda_0|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y})dF_{\mathbf{d},\mathbf{c}}(\mathbf{d},\mathbf{c}),\ \text{then it must be that}$$

$$\mathbf{1}\left\{\lambda_0\geq\arg\min_{x>0}\mathcal{U}(x|\mathbf{y},\mathbf{X},\sigma^2)\right\}=\mathbf{P}(\lambda\geq\lambda_0|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y}),$$

therefore we can conclude that

$$(\lambda|\sigma^2,\mathbf{d},\mathbf{c},\mathbf{y})=\mathcal{U}(x|\mathbf{y},\mathbf{X},\sigma^2)\text{ almost surely.}$$

If instead of using the prior (3.11), we decide to use (3.12) or (3.13), the full conditional posteriors of the parameters would be:

$$\left[(\mathbf{d}^\intercal\ \mathbf{c}^\intercal)^\intercal\,|\lambda,\sigma^2\right] \sim N_{l+k}\left(\mu_{\mathbf{dc}},\mathbf{\Sigma_{dc}}\right),$$

$$\left[\sigma^2|\lambda,\mathbf{d},\mathbf{c}\right] \sim Inv-Gamma\left(A_\epsilon+\frac{1}{2}n,\left[B_\epsilon^{-1}+\frac{1}{2}\sum_{i=1}^{n}\left(y_i-\eta(\mathbf{x}_i)\right)^2\right]^{-1}\right),$$

and λ (and θ_i's) fixed as

$$(\lambda|\mathbf{y},\mathbf{X})=\arg\min_{x>0}\mathcal{V}(x|\mathbf{y},\mathbf{X},\alpha),\ \text{if prior (3.12) was used, or}$$

$$(\lambda|\mathbf{y},\mathbf{X})=\arg\min_{x>0}\mathcal{M}(x|\mathbf{y},\mathbf{X})\ \text{if prior (3.13) was used.}$$

3.2.2 Bayesian regression model using tensor thin plate splines

The model in this section has the same form as for the thin plate splines (Section 3.2.1) with the only differences observed on the penalty term (2.2) and the corresponding reproducing kernel R. Therefore, the Bayesian model is the same but the basis functions changes. We now describe the changes on the basis functions.

Lets consider the reproducing kernel $R_{J_m^d}$ of the thin plate spline minimization problem. By the arguments and interpretation provided in Section 2.1.2, the following expressions are reproducing kernels for the tensor thin plate spline setting in the case $\mathbb{X} = \mathbb{R}^2$, with $\mathbf{x} = (\mathbf{x}_{(1)}, \mathbf{x}_{(2)}) \in \mathbb{R}^2$,

$$R_{K_1}(\mathbf{x}, \mathbf{y}) = \theta_1 R_{J_m^1}\left(\mathbf{x}_{(1)}, \mathbf{y}_{(1)}\right) + \theta_2 R_{J_m^1}\left(\mathbf{x}_{(2)}, \mathbf{y}_{(2)}\right), \tag{3.14}$$

$$\begin{aligned} R_{K_2}(\mathbf{x}, \mathbf{y}) =& \theta_1 R_{J_m^1}\left(\mathbf{x}_{(1)}, \mathbf{y}_{(1)}\right) + \theta_2 R_{J_m^1}\left(\mathbf{x}_{(2)}, \mathbf{y}_{(2)}\right) \\ &+ \theta_3 R_{J_m^1}\left(\mathbf{x}_{(1)}, \mathbf{y}_{(1)}\right) R_{0,m}\left(\mathbf{x}_{(1)}, \mathbf{y}_{(1)}\right) + \theta_4 R_{J_m^1}\left(\mathbf{x}_{(2)}, \mathbf{y}_{(2)}\right) R_{0,m}\left(\mathbf{x}_{(2)}, \mathbf{y}_{(2)}\right) \\ &+ \theta_5 R_{J_m^1}\left(\mathbf{x}_{(1)}, \mathbf{y}_{(1)}\right) R_{J_m^1}\left(\mathbf{x}_{(2)}, \mathbf{y}_{(2)}\right), \end{aligned} \tag{3.15}$$

where $R_{J_m^1}$ is the reproducing kernel for the thin plate splines in the domain $\mathbb{X} = \mathbb{R}$ described by (2.30), and $R_{0,m}$ is the reproducing kernel of the space of polynomials in $\mathbb{R}$ with degree smaller than $m + 1$; note $R_{0,m}$ is fully described by (2.28). The reproducing kernels (3.14) and (3.15) are the respective kernels for the tensor thin plate spline without interaction in the Anova decomposition of the associated Hilbert space, (Aronszajn (1950); Akhiezer and Glazman (1981a,b); Gu (2013)), while the setting with interaction terms has reproducing kernel (3.15).

For the simulation study in Section 3.3 we use the tensor thin plate spline with interaction, hence the reproducing kernel (3.15). In this case we need to choose the smoothing parameters $\{\theta_i\}_{i=1}^p$ and set $\lambda = 1$ for identificability reasons. The Bayes model interpretation has the same form as in the thin plate spline, but now the matrices Q, R and the projection matrix $A(\lambda)$ in (2.21) – (2.22) depend on $\{\theta_i\}_{i=1}^p$ with $\lambda = 1$; the functions $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{M}$ depend as well on $\{\theta_i\}_{i=1}^p$ and the priors (3.11), (3.12), (3.13) are specified minimizing over the positive quadrant of $\mathbb{R}^p$ ($p = 5$ here).

3.2.3 Full Bayes linear mixed effects model

The full Bayes model follows directly from the linear mixed model in Section 2.2.2. Here, we describe the priors on the parameters chosen such that the mean of the full conditional posterior of the fixed coefficients and predictors of the random effect $(\mathbf{d}^\intercal, \mathbf{c}^\intercal)^\intercal$ is expression (2.53). Consider the linear mixed model

$$\begin{aligned} &\mathbf{y}|\mathbf{d},\, \mathbf{c},\, \mathbf{e} = S\mathbf{d} + U\mathbf{c} + \mathbf{e}, \\ &\begin{pmatrix} \mathbf{c} \\ \mathbf{e} \end{pmatrix} \sim \mathbf{N}_{n+k}\left(\mathbf{0}, \begin{pmatrix} \sigma_c^2 I_k & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 I_n \end{pmatrix}\right), \end{aligned} \tag{3.16}$$

with $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ as before and $U = FV^{-1/2}$, $F_{i,j} = E_{m,d}\|\mathbf{x}_i - \mathbf{z}_j\|$, $V_{i,j} = E_{m,d}\|\mathbf{z}_i - \mathbf{z}_j\|$. Consider the priors

$$\begin{aligned} \mathbf{d} &\sim 1 \\ \sigma_c^2 &\sim Inv - Gamma\,(A_c, B_c) \\ \sigma_e^2 &\sim Inv - Gamma\,(A_e, B_e)\,. \end{aligned}$$

That the posterior distribution of the parameters is proper follows by similar arguments as in Proposition 74. The full conditional posterior distributions are

$$\begin{aligned} \left[(\mathbf{d}^\intercal\, \mathbf{c}^\intercal)^\intercal | \sigma_c^2, \sigma_e^2\right] &\sim N_{k+l}\left(\left([V\,U]^\intercal\,[V\,U] + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{D}\right)^{-1}[V\,U]\,\mathbf{y}, \left([V\,U]^\intercal\,[V\,U] + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{D}\right)^{-1}\right) \\ \left[\sigma_c^2 | \sigma_e^2, \mathbf{d}, \mathbf{c}\right] &\sim Inv - Gamma\left(A_c + k/2, \left(B_c + \frac{1}{2}\|\mathbf{c}\|^2\right)^{-1}\right) \\ \left[\sigma_e^2 | \sigma_c^2, \mathbf{d}, \mathbf{c}\right] &\sim Inv - Gamma\left(A_e + n/2, \left(B_e + \frac{1}{2}\|\mathbf{y} - S\mathbf{d} - U\mathbf{c}\|^2\right)^{-1}\right), \end{aligned}$$

where $\mathbf{D} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_l \end{pmatrix}$. Observe that the ratio σ_e^2/σ_c^2 plays the role of $n\lambda$, the smoothing parameter. It could be interpreted as assigning a prior to λ and observing the corresponding distribution on $\sigma_c^2 = \frac{\sigma_e^2}{n\lambda}$ which in this case is an inverse gamma. The next models use this approach, a prior is assigned to λ and σ_e^2, and σ_c^2 follows the corresponding induced prior.

The Bayesian model from this section was inspired by the full Bayes model to estimate function in the presence of errors in covariance by (Berry et al. (2002)).

3.2.4 Bayesian linear mixed model interpretation and empirical bandwidth parameters

Consider the linear mixed model (3.16) and define $\lambda = \frac{\sigma_e^2}{n\sigma_c^2}$. Consider the priors

$$\begin{aligned}
\mathbf{d} &\sim 1, \\
\lambda|\sigma_e^2 &= \arg\min_{x>0} \left\{\mathcal{U}(x|\sigma_e^2)\right\} \text{ almost surely}, \\
\sigma_e^2 &\sim Inv - Gamma\left(A_e, B_e\right).
\end{aligned}$$

The full conditional posterior of the parameters are

$$\begin{aligned}
\left[(\mathbf{d}^\intercal\, \mathbf{c}^\intercal)^\intercal|\lambda,\sigma_e^2\right] &\sim N_{k+l}\left(\left([V\,U]^\intercal\,[V\,U] + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{D}\right)^{-1}[V\,U]\,\mathbf{y}, \left([V\,U]^\intercal\,[V\,U] + \frac{\sigma_e^2}{\sigma_c^2}\mathbf{D}\right)^{-1}\right) \\
\lambda|\sigma_e^2 &= \arg\min_{x>0} \left\{\mathcal{U}(x|\sigma_e^2)\right\} \text{ almost surely}, \\
\left[\sigma_e^2|\lambda,\mathbf{d},\mathbf{c}\right] &\sim Inv - Gamma\left(A_e + n/2, \left(B_e + \frac{1}{2}\|\mathbf{y} - S\mathbf{d} - U\mathbf{c}\|^2\right)^{-1}\right),
\end{aligned}$$

Alternatively, in two different models, λ can be chosen as well using (3.12) or (3.13) and the full conditional posteriors of the parameters would be similar to above but λ as the *argmin* of the respective function.

3.3 Simulation Study: Bayesian Models using Thin Plate Splines, Tensor Thin Plate Splines and Linear Mixed Model Interpretation

We perform a simulation study to compare the performance of the models in Section 3.2 in terms of point estimates and coverage of credible intervals for functions $\eta : \mathbb{R}^2 \to \mathbb{R}$. The algorithm and our methods work in theory for any number of covariates but we provide a robust simulation study for the case of two covariates. We tested our methods with some examples to estimate functions with domain in $\mathbb{R}^3$ and $\mathbb{R}^4$ obtaining good resulting estimation. The purpose of the simulation study is to observe the performance on the estimation provided by all 10 models measured in terms of the Bayes estimates of the function, Bayes prediction, and empirical coverage of the credible intervals for predictions of η. Based on these results we choose an acceptable model in order to face the problem of nonparametric regression with errors in variables in Section 4.

There has been previous work on simulation studies for nonparametric regression models with Bayes interpretation, such in (Wahba (1983); Nychka (1988); Kim and Gu (2004)), but most of the studies are for estimation of univariate functions or examples of estimation for bivariate functions are simply shown. Wahba (1983) and Nychka (1988) report an average coverage probability across the region of estimation for the credible intervals that is similar to the level of the credibility, when using the GCV method to choose the bandwidth parameters. Nychka (1988) mentions that the main disadvantage of the approach is that the "confidence intervals" are only valid in an average sense over the region of estimation, and may not be reliable if evaluated for pointwise estimation or only evaluated at peaks or troughs in the estimate; the pointwise coverage of the credible intervals depend on the unknown function. With the simulation study we have designed, we compare four methods to choose the smoothing parameters. We can observe the impact of having more than one smoothing parameters in the model (tensor thin plate splines) or choosing the smoothing parameter by assigning a prior. We are able to observe the behavior of the point estimates and the empirical coverage of credible intervals in different regions of estimation varying in size and the concentration of observed data for bivariate function.

Table 3.1 Summary form of the models and smoothing methods to choose/estimate the smoothing parameters in the simulation study. LMM - linear mixed model interpretation, TPS - thin plate splines, TTPS - tensor thin plate splines with anova interaction. UERL - unbiased estimate relative loss, GCV - generalized cross validation, REML - restricted maximum likelihood under the Bayes model, Bayes - inverse gamma prior on both variances σ_e^2 and σ_c^2.

Smoothing Model	**Bandwidth Method**
LMM	UERL
LMM	GCV
LMM	RML
LMM	Bayes
TPS	UERL
TPS	GCV
TPS	RML
TTPS	UERL
TTPS	GCV
TTPS	RML

The form of the competing models and the methods to select the bandwidth parameters are summarized in Table 3.3. There, we use the abbreviations LMM for the linear mixed models described in Sections (3.2.3) and (3.2.4), TPS for the Bayesian models using thin plate splines (Section (3.2.1)), and TTPS for the Bayesian models using tensor thin plate splines (Section (3.2.2)). With regard to the selection of the smoothing parameters, the notation UERL indicates that λ (and θ_i's) were assigned with the prior (3.11), GCV indicates that the prior (3.12) was used, and a Bayesian model with (3.13) as prior is denoted as RML. In the same table, the combination LMM and *Bayes* bandwidth method denotes the model described in Section (3.2.3), and rest of the LMM models were described in Section (3.2.4). The former model will denoted as *Full Bayes* model, and the later models will be jointly denoted as *Bayes Empirical* models.

For each combination of the values of the parameters, 200 simulated data sets were generated, with $\{\mathbf{x}_i\}_{i=1}^n, \mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$, $n = 50,\ 100,\ 400,$ and 800; the responses $\{y_i\}_{i=1}^n$ were simulated with the form $y_i = \eta(\mathbf{x}_i) + \epsilon_i$, while $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ using $\sigma = 0.01,\ 0.1,\ 0.25,\ 0.5,$ and 1. The deterministic function to be estimated is

$$\eta((x_{(1)}, x_{(2)})) = \left[4 + \frac{\sin(\frac{1}{2}\pi x_{(1)})}{1 + 4x_{(1)}^2 \mathbf{1}(x_{(1)} > 0)}\right] \times \left[\sin(x_{(2)}) + \cos(x_{(2)}) + x_{(2)}\right]. \tag{3.17}$$

For each of the simulations, the priors on the variances were chosen to be $\sigma_\epsilon^2 \sim Inv - Gamma(1,1)$, $\sigma_c^2 \sim Inv - Gamma(1,1)$, and $\sigma_e^2 \sim Inv - Gamma(1,1)$ for the respective models. The values of the hyper-parameters for the priors were chosen in this way because the prior information they provide is not strong. It was found that estimation is insensitive to moderate modifications to these values.

Estimates for $\eta(\chi)$ for any $\chi \in \mathbb{R}^2$ are found as the mean of the posterior predictive distribution, (Gelman et al., 2014, pag 7):

$$\Pi(\eta(\chi_i)|\mathbf{y}) = \int \Pi\left(\eta(\chi)|\theta\right) \Pi\left(\theta|\mathbf{y}\right) d\theta.$$

Observe that we are abusing of the notation, η is a deterministic functions while the notation $\Pi(\eta|\mathbf{Y})$ suggest that η is a random process with variance function, in principle, non zero. We use this notation keeping in mind that we are interested in the posterior mean of $\Pi(\eta(\chi_i)|\mathbf{Y}) = [\eta(\chi_i)|\mathbf{Y}]$ as point estimate of the deterministic $\eta(\chi_i)$.

The posterior distribution of the parameters in the models and the posterior predictive distribution $\Pi(\eta(\chi)|\mathbf{Y})$ do not have an analytically form. Samples were drawn using MCMC methods. Two independent chains with different initial overdispersed values for each parameter were drawn using Gibbs sampler, (Gelman et al., 2014, pag 276 - 278). Each of the chains were run for $10,000$ iterations discarding the first $7,000$ realizations as burn in and thinning the rest of the sequences by keeping every 3 draws. In the simulation study is not possible to assess convergence of the MCMC chains for all simulated parameters and all data sets at the same time. Instead, convergence tests such as Geweke test, (Geweke et al. (1991)) and Gelman test (Gelman et al., 2014, pag. 285) were used to separately test the convergence for the parameters.

Realizations of the posterior predictive distribution are achieved using the samples from the posteriors. Let $\hat{\eta}(\chi) := \mathbb{E}\left[\Pi(\eta(\chi)|\mathbf{Y})\right]$ denote the point estimator of $\eta(\chi)$ and $\tilde{\eta}(\chi) := \mathrm{sd}\left[\Pi(\eta(\chi)|\mathbf{Y})\right]$ denote the standard deviation of the posterior distribution. Then $\hat{\eta}(\chi)$ and $\tilde{\eta}(\chi)$ are approximated with the sample mean and the unbiased sample standard deviation from the realizations of the posterior predictive distributions.

Figure 3.1 shows an example of estimation using the model obtained from the Bayes interpretation of the thin plate splines in a grid of resolution 0.05×0.05 inside the square $[-2.25, 2.25]^2$. We denote this grid as $\{\chi_i\}_{i=1}^N \subset \mathbb{R}^2$. The square $[-2.25, 2.25]^2$ was chosen as the region of estimation for all the simulated data sets of the study because it has the property that it would contain about 95.17% of all points simulated from $N_2(\mathbf{0}, I_2)$; the grid was not chosen to be finer because of storage availability. The Bayes point estimator $\hat{\eta}$ is expected to behave similarly to frequentist estimation of the non-parametric regression obtained by (1.1) because of the interpretation of the mean of the full conditional posterior distribution as a solution. Observe that the pointwise standard deviation in the estimation $\tilde{\eta}$ is larger on the boundaries of the region of estimation because have less information about the regression function in that area, thought the standard deviation is smaller in the center of the region as expected.

Each model in Table 3.3 does not assume a parametric form for η, instead it is assumed that η is in the space $\mathcal{H}^\star$ (2.48); in this way we approximate the solution to (1.1) in $\mathcal{H} \supseteq \mathcal{H}^\star$, the *RKHS* generated by the reproducing kernel (2.30). Such a *RKHS* , for the estimation in Figure 3.1, does not contain functions that are not at least three ($m = 3$) times partially

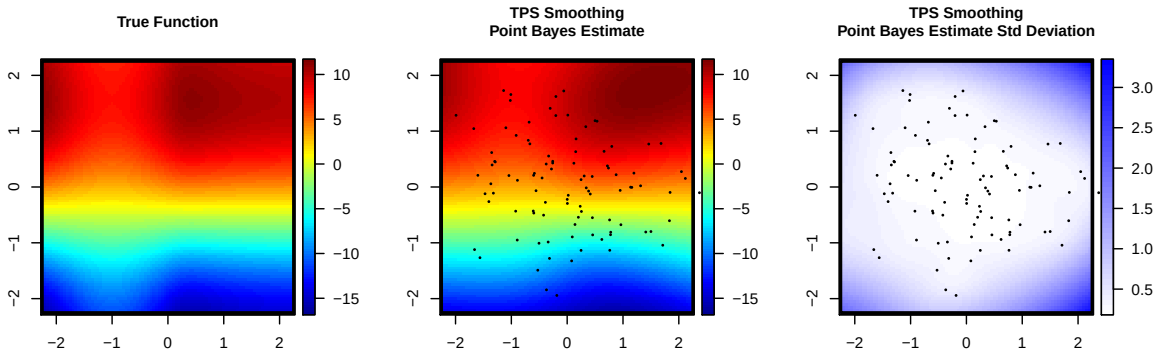

Figure 3.1 Level curves for the true function η equation (3.17) (left), point Bayes estimate $\hat{\eta}(\chi)$ (center), and pointwise standard deviation, $\tilde{\eta}(\chi)$ (right), for estimation using Bayes model interpretation of the thin plate splines with $m = 3$ and smoothing parameter chosen using the restricted maximum likelihood method, with $n = 100$ and $\sigma^2 = 0.5$. The dots in the plots are the observed values of the covariates $\{\mathbf{x}_i\}_{i=1}^{100}$.

differentiable where the squares of the partial derivatives of degree $m = 3$ are integrable. Therefore the strongest assumptions made on η is that all of its third partial derivatives exist for every point in $\mathbb{R}^2$ and the integral over $\mathbb{R}^2$ of the square of each of the derivatives is finite. It is possible to impose a weaker assumption on η and the proposed methodology would still be theoretically justified and interpretations would be the same; it would be enough to have that all the partial derivatives of degree $m = 3$ are integrable in the square $[-2.25, 2.25]^2$, but we did not pursuit to prove this statement. The required properties of partial differentiabilty, integrability of the square of the partial derivatives, and that the mean of the full conditional distribution of η is the function that best interpolates the data as measured by the squared loss function $\frac{1}{n}\sum_{i=1}^{n}(y_i - \eta(\mathbf{x}_i))^2$ subject to the constrain that $J_3^2(\eta)$ is small (Theorem 51), are the assumptions that lead us to compute $\hat{\eta}$ as observed in Figure 3.1.

3.3.1 Prediction and variability of prediction for the target regression function and discussion

We propose a summary to evaluate the performance of the estimated functions $\hat{\eta}$ as follow. By the interpretation of the mean of the full conditional posterior of the models as thin plate splines or tensor thin plate splines, and by the way the smoothing parameters were chosen

using the Unbiased Estimate Relative Loss method (2.55) and the generalized cross validation method (2.59), $\hat{\eta}$ approximately minimizes the loss function $\frac{1}{n}\sum_{i=1}^{n}(\eta_\lambda(\mathbf{x}_i)-\eta(\mathbf{x}_i))^2$, (2.54) up to the constrain described in Theorem 51. Instead of evaluating the performance of the estimation $\hat{\eta}$ using the loss function $N^{-1}\sum_{i=1}^{N}(y_i-\eta(\mathbf{x}_i))^2$ (as we know η_λ already minimizes a form like this), we propose to use a summary that is function of absolute differences: $|\eta(\mathbf{x})-\hat{\eta}(\mathbf{x})|$. For consistency we use this summary for all models. The mean absolute error $N^{-1}\sum_{i=1}^{N}|y_i-\eta(\mathbf{x}_i)|$ is a robust measure of predictive accuracy, it tends to prefer predictions procedures that on average are reasonably good and is less sensitive to large deviations than the square loss function.We propose the use of

$$\frac{1}{N\times\sigma}\sum_{i=1}^{N}|\eta(\chi_i)-[\eta(\chi_i)|\,\mathbf{Y}]| \tag{3.18}$$

to compute and compare deviations from the real function and to measure variability of the predictions around η for all models in Table 3.3. As the previous expression is a random variable, we approximate the mean

$$MAPE := \frac{1}{N\times\sigma}\mathbb{E}\left[\sum_{i=1}^{N}|\eta(\chi_i)-[\eta(\chi_i)|\,\mathbf{Y}]|\right] \tag{3.19}$$

and its respective variability

$$SDMAPE := \frac{1}{N\times\sigma}\sqrt{\mathbf{Var}\left[\sum_{i=1}^{N}|\eta(\chi_i)-[\eta(\chi_i)|\,\mathbf{Y}]|\right]}. \tag{3.20}$$

for each combination of parameters of the simulated data and each of the 200 repetitions. Above, MAPE stands for *Mean Absolute Prediction Error*. Estimation of (3.19) and (3.20) is achieved using the realizations of $[\eta(\chi_i)|\,\mathbf{Y}]$, computing (3.18) and finally obtaining the sample mean and sample standard deviation.

Table 3.3.1 summarizes a part of the results for the computations of $MAPE$ and $SDMAPE$. This table summarizes the case for $n = 100$ and $\sigma^2 = 0.5$. Column **Avg M** is the average across 200 computed $MAPE$ for each model, **M 25** and **M 75** are the 25% and 75% empirical percentiles of these 200 computed $MAPE$. **Avg SDM** is the average of the 200 computed $SDMAPE$ with their respective 25% and 75% percentiles. Bold number for column **M 25** indicates that the 25% percentile computed is larger than at least one of the 75% percentile

Table 3.2 Part summary simulation results for $MAPE$ and $SDMAPE$. Simulated data with link function (3.17), $n = 100$ and $\sigma^2 = 0.5$. Using 200 repetitions, **Avg M** is the average of the computed $MAPE$, **M 25** and **M 75** are the 25% and 75% empirical percentile of the 200 computed $MAPE$. **Avg SDM** is the average of the 200 computed $SDMAPE$ with their respective 25% and 75% percentiles. Black number for column **M 25** indicate that the 25% percentiles computed with the respective model is larger than at least one of the 75% $MAPE$ percentile computed for another model. Black numbers for column **M 75** indicates that there is at least one 25% percentile $MAPE$ from another model that is larger than this 75% percentile. Similarly for columns **SDM 25** and **SDM 75**.

Model	**m**	**Avg M**	**M 25**	**M 75**	**Avg SDM**	**SDM 25**	**SDM 75**
Bayes Empcal GCV	2	0.96	**0.86**	**1.04**	0.10	0.09	**0.1**
Bayes Empcal GCV	3	0.96	**0.83**	**1.09**	0.14	**0.11**	**0.16**
Bayes Empcal GCV	4	1.06	**0.9**	**1.14**	0.19	**0.15**	**0.22**
Bayes Empcal RML	2	0.94	**0.84**	**1.03**	0.10	0.09	**0.11**
Bayes Empcal RML	3	0.96	**0.82**	**1.08**	0.15	**0.12**	**0.17**
Bayes Empcal RML	4	1.07	**0.92**	**1.14**	0.20	**0.16**	**0.23**
Bayes Empcal UERL	2	0.94	**0.85**	**1.02**	0.10	0.09	**0.11**
Bayes Empcal UERL	3	0.96	**0.82**	**1.08**	0.15	**0.12**	**0.17**
Bayes Empcal UERL	4	1.11	**0.95**	**1.2**	0.22	**0.16**	**0.24**
Full Bayes	2	0.94	**0.84**	**1.03**	0.10	0.09	**0.11**
Full Bayes	3	1.12	**0.91**	**1.28**	0.17	**0.11**	**0.17**
Full Bayes	4	1.51	**1.34**	1.59	0.15	**0.11**	**0.17**
Tensor TPS GCV	2	0.65	0.55	**0.74**	0.60	**0.55**	**0.65**
Tensor TPS GCV	3	0.87	0.67	**0.98**	0.84	**0.7**	**0.94**
Tensor TPS GCV	4	1.37	**0.92**	1.69	1.30	**0.99**	1.48
Tensor TPS RML	2	0.65	0.53	**0.72**	0.66	**0.6**	**0.71**
Tensor TPS RML	3	0.84	0.62	**0.96**	0.90	**0.74**	**1**
Tensor TPS RML	4	1.39	**0.91**	1.76	1.41	**1.08**	1.59
Tensor TPS UERL	2	0.66	0.56	**0.76**	0.62	**0.57**	**0.66**
Tensor TPS UERL	3	0.88	0.68	**0.98**	0.87	**0.74**	**0.95**
Tensor TPS UERL	4	1.43	**0.91**	1.77	1.34	**1.02**	1.51
TPS GCV	2	0.81	0.72	**0.93**	0.54	**0.51**	**0.56**
TPS GCV	3	0.78	0.64	**0.9**	0.66	**0.61**	**0.71**
TPS GCV	4	0.81	0.66	**0.95**	0.81	**0.7**	**0.88**
TPS RML	2	0.77	0.67	**0.86**	0.58	**0.55**	**0.61**
TPS RML	3	0.75	0.61	**0.87**	0.72	**0.66**	**0.77**
TPS RML	4	0.80	0.65	**0.94**	0.89	**0.77**	**0.95**
TPS UERL	2	0.78	0.69	**0.88**	0.56	**0.54**	**0.58**
TPS UERL	3	0.74	0.61	**0.86**	0.73	**0.66**	**0.77**
TPS UERL	4	0.81	0.65	**0.94**	0.96	**0.82**	**1.04**

computed for another model; the graphical interpretation of boxplots in Figure 3.2 is that boxes of these cases do not intersect. Bold numbers for column **M 75** indicate that there is at least one 25% percentile from another model that is larger than this 75% percentile for this specific combination of parameters $n = 100$ and $\sigma^2 = 0.5$. Complete graphical summaries for columns **Avg M**, **M 25** and **M 75** are presented in Figure *C*.1. Similar graphical representations for the last three columns of Table 3.3.1 are shown in Figures 3.3 and *C*.2 in Appendix *C*.

The striking features of this Table and its graphical displays (Figures 3.2, *C*.1) for the $MAPE$ columns is that the frequentist properties for the estimates of η as measured by (3.19) seems to be similar within TTPS, TPS, and *Empirical Bayes* models regardless of the score minimization criteria used to choose the smoothing parameters, with the exception of the *Bayes* bandwidth method (Table (3.3)). The TPS models have as good estimates as the TTPS models when the variance σ^2 is not small ($\sigma^2 \neq 0.1^2$) and when comparing across different values of m. This suggests at least for this example, that one smoothing parameter λ is enough to provide similar estimates of η as when using five smoothing parameters $\{\theta_i\}_{i=1}^5$ in the TTPS models.

The *Full Bayes* model performs in a similar way to the *Bayes Empirical* models in the sense that there is not practical difference in the median of the MAPE for the 200 repetitions when comparing across values of the parameter m for the same σ^2. For the *Full Bayes* model and the *Bayes Empirical*, we did not find practical differences for predicting in the square $[-2.25, 2.25]^2$. Similar conclusions hold comparing the predictions of the *Bayes Empirical* and TTPS or with the TPS models. For now, we leave the comparison of the MAPE between the *Full Bayes* model and the TTPS and TPS models until after we discuss the SDMAPE for the *Bayes Empirical* models, TTPS and TPS.

The variability of the marginal posterior process $[\eta|\mathbf{y}]$ around η is positively related with the $SDMAPE$ and discussed now. The most striking feature in the last three columns of Table 3.3.1 and represented in Figure 3.3 and *C*.2, is that the average variability of the predictions of η over the grid in the square $[-2.25, 2.25]^2$ is not statistically significant or different between methods to choose the smoothing parameters and within the models TTPS, TPS or Bayes Empirical. For the cases shown in Table 3.3.1 and Figure 3.3, there is no statistically significant difference in the average variability around η of the predictions in the grid for the TTPS and

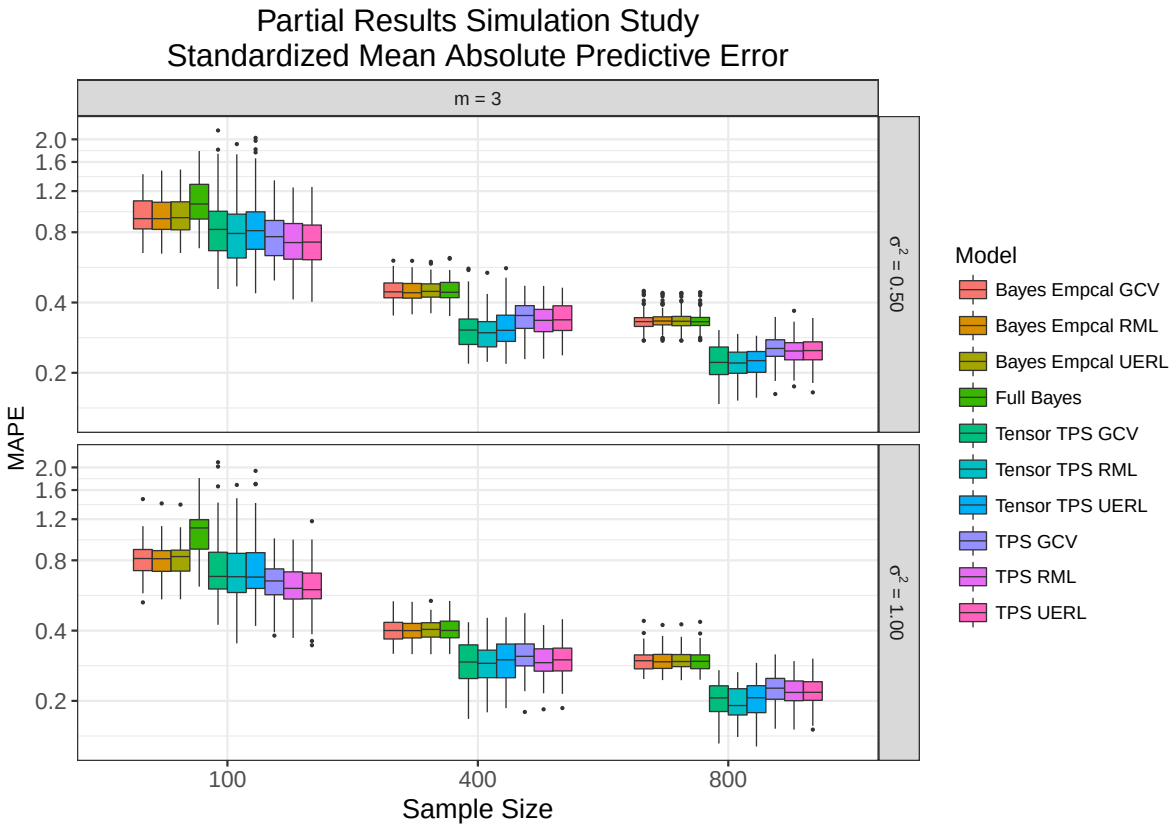

Figure 3.2 Boxplots part of the simulation results for the standardized mean absolute predictive error (3.19) (MAPE) for the multivariate regression problem predicting over the grid of resolution 0.05×0.05 inside the square $[-2.25, 2.25]^2$. Observe that the y-axis is in the log_{10} scale. The rows indicates the true observation-error variance σ^2. The columns indicate the degree of derivative m for the penalty in (2.23). The models are described in Table 3.3. Complete simulation results appear in Figure C.1.

TPS models (median of SDMAPE for 200 repetitions). For the few cases when there is a statistical difference in the values of SDMAPE as seen in the appendix Figure C.2, that is to say, for the cases when the predictions from the TPS have smaller variability than those from the TTPS model as measured by the SDMAPE, the $MAPE$ is statistically smaller for TTPS than for TPS.

Based on the results of the point predictions for η and their variability around the true link function η, we have shown evidence at least for this simulated setting that the TTPS and TPS models are equally competitive, at least in a practical sense of predicting and extrapolating in the square $[-2.25, 2.25]^2$. For now, the advantage of TPS models over TTPS is that less

computational effort is required to estimate the single bandwidth parameter λ for the TPS model than the five bandwidth parameters $\{\theta_i\}_{i=1}^5$ for the TTPS model.

We admit that a function f can always be constructed such that a TTPS model with five smoothing parameters is required over the TPS model but, in similar way, another artificial function g can be constructed such that the TTPS over fits the data. The construction would follow from the criticism of Barry et al. (1986) that is mentioned in Section 2.1.1. For the example function η (3.17) that was chosen without being influenced a priori by the form of the models and that we use in this simulation study, it seems that both TPS and TTPS behave practically equal when predicting.

Now that we have discussed the SDMAPE, we come back to the MAPE summary for the *Full Bayes* model. The median of the MAPE over the 200 repetitions of the *Full Bayes* model in comparison with the medians of the TTPS model or with the TPS model is statistically larger. However, the difference is not clearly practical different as we argue now. While the median value of the MAPE seems to be larger than with the TTPS or TPS models, the median of the SDMAPE for these same cases of the *Full Bayes* model are large indicating that the variability of the point predictions for the $FullBayes$ model for each fitting, is large around the true function. This large variability is observed because the *Full Bayes* model produce larger pointwise credible intervals for $[\eta(\chi_i)|\mathbf{y}]$ than any other model. In order to evaluate if the credible intervals are too large we evaluate the empirical coverage of the credible intervals produced by each model in Section 3.3.3.

Recall the interpretation of λ in the *Full Bayes* model, from Section 3.2.3, as $n\lambda = \frac{\sigma_e^2}{\sigma_c^2}$. It is possible that by providing a more informative prior over σ_c^2 or σ_e^2 or even changing the family of the distribution of the priors for these variance parameters, the bias and the variability of the predictions for η decrease. We did not pursue this objective.

3.3.2 Observation error variance summary results and discussion

The main objective of the models evaluated in this dissertation work is to predict reasonably well evaluations of the target function. But the mechanism of estimating the link function

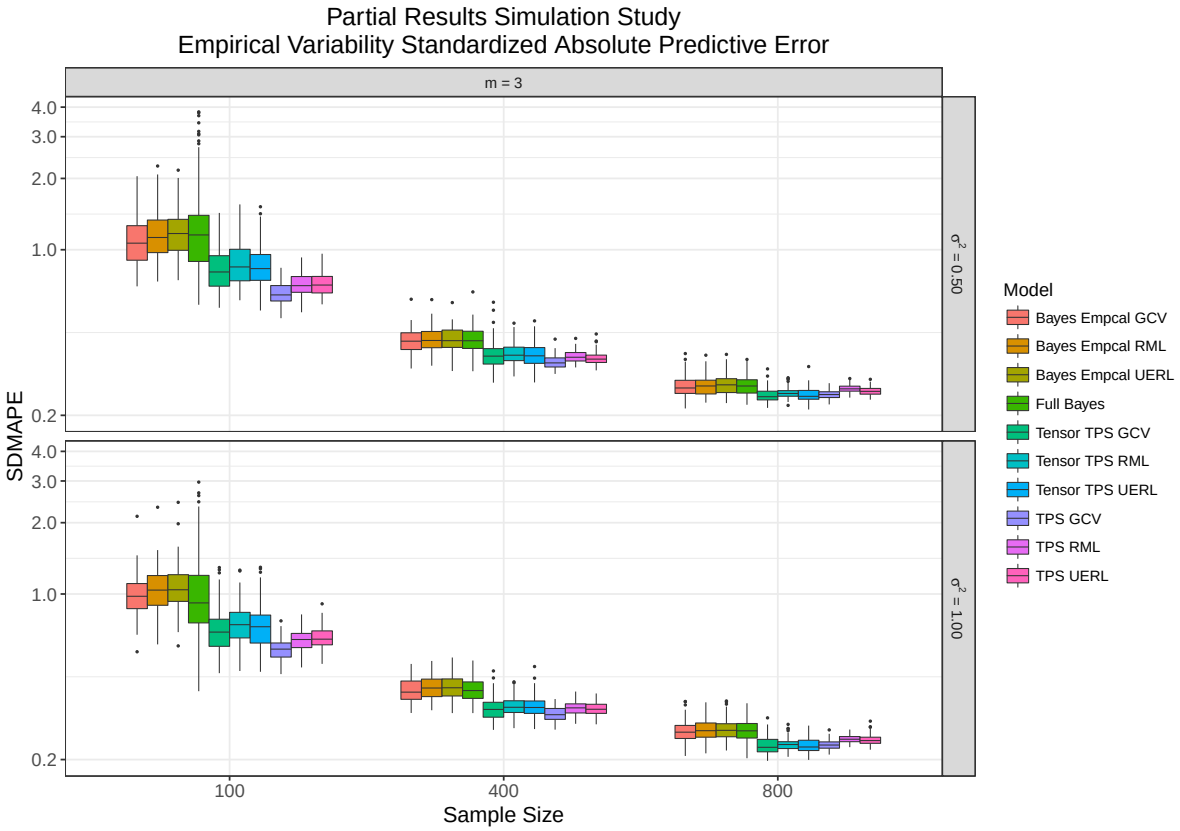

Figure 3.3 Boxplots part of the simulation results for the standardized standard deviation of the absolute predictive error (3.20) (SDMAPE) for the multivariate regression problem predicting over the grid of resolution 0.05×0.05 inside the square $[-2.25, 2.25]^2$. Observe that the y-axis is in the log_{10} scale. The rows indicates the true observation-error variance σ^2. The columns indicate the degree of derivative m for the penalty in (2.23). The models are described in Table 3.3. Complete simulation results appear in Figure C.2.

requires the estimation of the parameters of the variance σ^2 of the errors $\{\epsilon\}_{i=1}^n$. We summarize in this section the means of the marginal posterior distributions of σ^2.

Figure 3.4 present the boxplots of the means of the marginal posterior distributions for σ^2. Each boxplot is computed from the 200 simulated data sets and each combination of parameters and model fitting. The vertical axis is the ratio between the estimated mean and the true value of the variance. Although Figure 3.4 shown that only the *Full Bayes* models and the *Bayes Empirical GCV* produce biased estimation of the variance, the more comprehensive Figure C.3 reveals that in about half of the combination of parameters, the *Bayes Empirical* models over-estimate the variance by a factor of 2 or 3 and in a few extreme cases ($\sigma^2 = 0.1^2$),

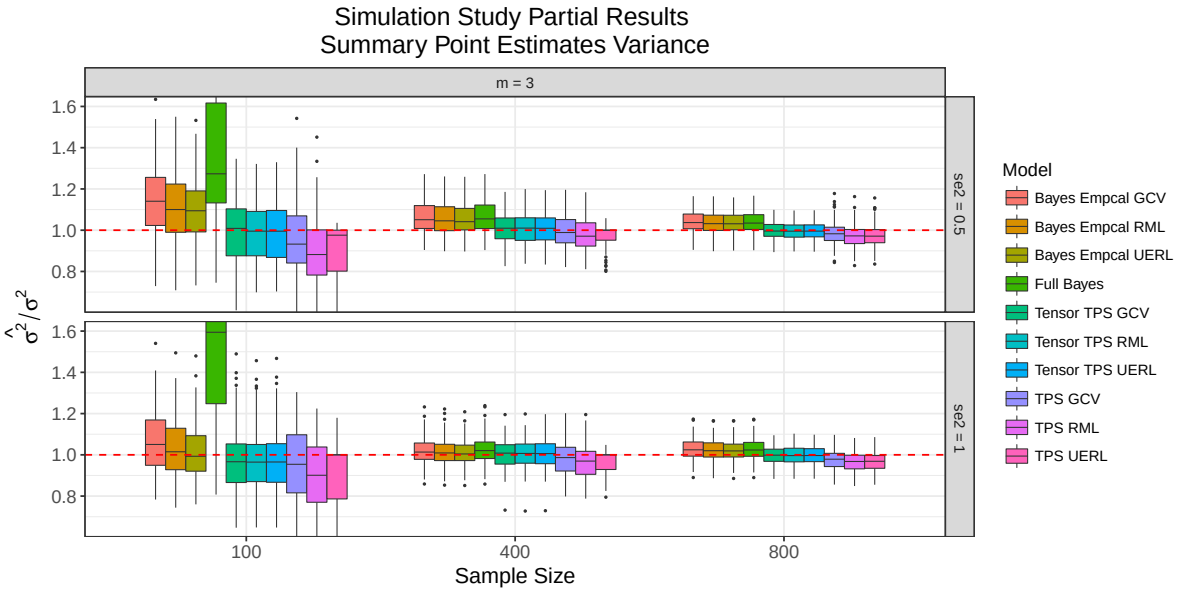

Figure 3.4 Mean marginal posterior of the observation-error variance parameter σ^2. Each column indicate the true observation-error variance σ^2. The models are described in Table 3.3. Observe that in the $y - axis$ is plotted the ratio of the estimated variance and the true variance; it is desired to have a ratio of 1, furthermore, the y-axis is in the log_{10} scale and each plot has a different range of values.

by a factor of more than 10. The *Full Bayes* model produces more extreme biased estimator of the variance of the errors. For these models, the credible intervals for σ^2 obtained using the marginal posterior distribution $\left[\sigma^2|\mathbf{y}\right]$ are large and sometimes capture the true variance parameter even in the extreme cases of overestimation. But, of course, large credible intervals are not desired. We do not show any plot summary for the variance of credible intervals.

The TTPS models estimate extremely well the variance parameter σ^2 through the mean of the marginal posterior predictive distribution. The boxplots in Figures 3.4 and *C*.3 contain always the value 1 even for two out of the three most difficult cases for estimating $\sigma^2 = 0.1^2$. Even when there seems to be bias in the estimation of $\sigma^2 = 0.1^2$ with $m = 2$, the pointwise credible intervals generated from $\left[\sigma^2|\mathbf{y}\right]$ would capture 0.1^2 with an acceptable credibility level.

The TPS models overestimate the variance parameter in few cases, specially the case $m = 2$ and $\sigma^2 = 0.1^2$. The over estimation is by a factor of 4 in the most extreme case but the rest of the overestimation is at most by a factor of 1.5. In all cases, the pointwise credible intervals for each repetition, case, and method to choose the smoothing parameters, covers the respective

value of σ^2 with the corresponding credibility level. We do not shown further summary of the coverage of the credible intervals in this Section.

Clearly, the TTPS models produce the best estimation of the variance parameter from the perspective of point estimates using the mean of the marginal posterior distribution $\left[\sigma^2|\mathbf{y}\right]$, but the TPS model also produces acceptable estimation in terms of pointwise credible intervals.

3.3.3 Empirical coverage of credible intervals from predictive posterior distribution for the target regression function and discussion

We evaluate now the empirical pointwise coverage of the credible intervals for $\eta(\chi_i)$ for all χ_i in the grid and C-level pointwise credible intervals for each $\eta(\chi_i)$. For each independent simulation $j = 1, \cdots, 200$ define the variable

$$\xi_i^{(j)} = \begin{cases} 1 & \text{if } \eta(\chi_i) \text{ is contained in the } C\% \text{ centered credible intval of } \Pi(\eta(\chi_i)|\mathbf{Y}) \text{ from simulation } j, \\ 0 & \text{if } \eta(\chi_i) \text{ is not contained in the } C\% \text{ centered credible intval of } \Pi(\eta(\chi_i)|\mathbf{Y}) \text{ from simulation } j. \end{cases}$$

The random variables $\xi_i^{(j)} \sim Bernoulli(\rho_i)$ are Bernoulli distributed, and the random variables ρ_i have the same expected value $\zeta \in [0, 1]$. If the points $\{\chi_i\}_{i=1}^N$ where we try to predict η are the observation points $\{\mathbf{x}_i\}_{i=1}^N$ then the defined ζ is known as the *average coverage probability* (ACP) (Wahba (1983)). Wahba estimates ACP using a single data set ($j = 1$) for the $C\%$ level using the expression

$$\hat{\zeta} = ACP(C) := \frac{1}{n} \# \left\{ i : |\eta(\mathbf{x}_i) - \hat{\eta}_{\lambda_v}(\mathbf{x}_i)| < z_{\alpha/2} \sigma_v^2 \sqrt{[A(\lambda_v)]_{ii}} \right\} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^{(1)}, \tag{3.21}$$

where $z_{\alpha/2}$ is the $1 - \alpha/2$ quartile of the standard normal distribution, $\alpha = 1 - C/100$. The ACP(C) (3.21) is obtained from a centered credible interval from Wahba's Bayesian model where the prior on σ^2 is degenerately σ_v^2 or equivalently $\sigma^2 = \sigma_v^2$ is assumed and $\lambda = \lambda_v$. We use the same definition of ACP for any grid $\{\chi_i\}_{i=1}^N$ and we estimate it using the credible intervals from the marginal posterior distributions $\{[\eta(\chi_i)|\mathbf{y}]\}_{i=1}^N$ or equivalently it can be estimated by obtaining the sample mean of $\{\rho_i\}_{i=1}^N$:

$$\begin{aligned}\hat{\zeta} = ACP(C) &= \frac{1}{N \times 200} \sum_{i=1}^{N} \sum_{j=1}^{200} \xi_i^{(j)} \\ &= \frac{1}{N} \sum_{i=1}^{N} \rho_i^{(j)}.\end{aligned} \tag{3.22}$$

It was investigated by Nychka (1988), in the setting of Wahba (1983) using large-sample approximation theory and simulation, that (3.21) is close to the nominal. Nychka (1988) reports *"From a frequency point of view, this agreement occurs because the average posterior variance for the spline is similar to a consistent estimate of the average squared error and because the average squared bias is a modest fraction of the total average squared error. These properties are independent of the Bayesian assumptions used to derive this confidence procedure"*. In our setting, we can only use the same arguments for the full conditional posterior distribution $[\eta|\sigma^2, \mathbf{y}, \lambda]$, as these arguments do not apply to the marginal posterior distribution and also not to the *Full Bayes* and *Bayes Empirical* models. However, part of the explanation about we obtaining similar results regarding to (3.22) being close to the nominal value for some models, as we will describe, must follow Nychka's arguments closely. Our purpose now is to evaluate the nominal level achieved by each of the ρ_i for each combination of parameters and models, and to test the nominal level achieved by ζ or ACP in our setting of bivariate regression with different algorithms to choose the bandwidth parameters.

A graphical display obtained from the computation of the ρ_i over the grid in the region $[-2.25, 2.25]$ using the 200 simulated data sets is shown in Figure 3.5 for the credibility levels 95%, 65% and 35%. A more comprehensive example using TPS model and RML for the bandwidth parameter is presented in Figure C.4. We would like that each ρ_i is close to the respective nominal level.

Observe in Figure 3.5 and in the appendix Figure C.4, that the nominal level for the ρ_i's seems to be approximately achieved in the center of the region of estimation while the empirical coverage decreases at the boundary of the regions of estimation. This effect is because we are extrapolating in extreme parts of the regions of estimation. We may evaluate the change in the empirical coverage distribution of the ρ_i's as we reduce the area of prediction of the function to

concentric smaller circles instead of on the square region $[-2.25, 2.25]$. The true function and an example of observed values $\{\mathbf{x}_i\}$ for Figure 3.5 are presented in Figure 3.1.

For each of the models in Table 3.3 and all combinations of the parameters $m \in \{2, 3, 4\}$, $\sigma^2 \in \{0.1^2, 0.1, 0.5^5, 0.5, 1\}$, we computed the empirical coverages ρ_i of 95% credible intervals from the respective marginal posteriors (left plot Figure 3.5). We summarize each plot as in Figure 3.5 or C.4 using boxplots of the ρ_i's. Each box in Figure 3.6 was computed from the respective ρ_i. A more comprehensive summary is shown in the appendix Figure C.5.

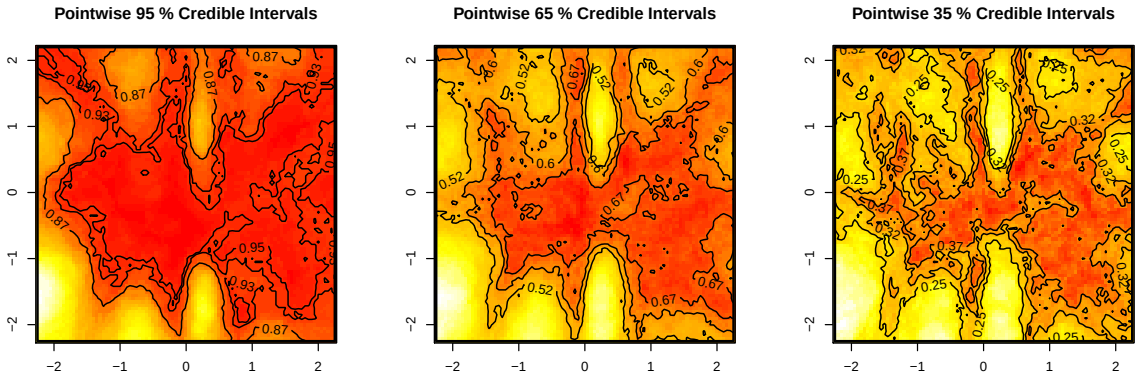

Figure 3.5 Level curves of the empirical coverage for the pointwise 95%, 65% and 35% credible intervals using the Bayesian model with thin plate splines, $m = 3$, and smoothing parameter λ chosen with the restricted maximum likelihood method. Each value in the level plot is an estimate of ρ_i for the mean coverage of $\eta(\chi_i)$ in the grid $\{\chi_i\}_{i=1}^{N}$ using 200 different simulated data sets and computing the respective credible intervals. Each data set was simulated with $n = 100$, $\sigma^2 = 0.5$ and $\mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$. The target function is described by (3.17) and plotted in Figure 3.1. Figure C.4 shows a more complete simulation for the credibility levels $C = 95\%$.

The first and most evident feature about Figure 3.6 is that for most of the models and methods to choose the smoothing parameters, $\hat{\zeta}$ is close to the nominal value 95% given that the boxes contain the value 0.95. The center 50% of the $\hat{\rho}_i$'s, in most cases, are within 20% points of the nominal value 0.95; or that the center 50% of the $\hat{\rho}_i$'s are in the interval $(0.78, 0.98)$. The few cases shown in this Figure are not enough to explain the distribution of pointwise coverages, so that we must observe the more comprehensive Figure C.5 in the appendix.

Let us first consider the cases when the simulated data was generated with $\sigma^2 > 0.1$. We can say that the ζ is close to the nominal value and the center 50% ρ_i's are fairly close the the

nominal value as well, around 20% points. For the rest of the ρ_i's, the ones contained in the whiskers, we have a diverse range of cases, some of them are about 30% points from the nominal level while some others have a empirical coverage as low as 25%. The ρ_i's that are considered as outliers could have a value as low as 10% or 0%; of course these ρ_i's correspond to the regions of estimation where we are extrapolating the function far away from the observed data. It may be surprising that the TTPS models provide fairly good pointwise empirical coverage, observe that the whiskers of the boxes for these cases are rarely below 50% and in most of the cases they are above 60%.

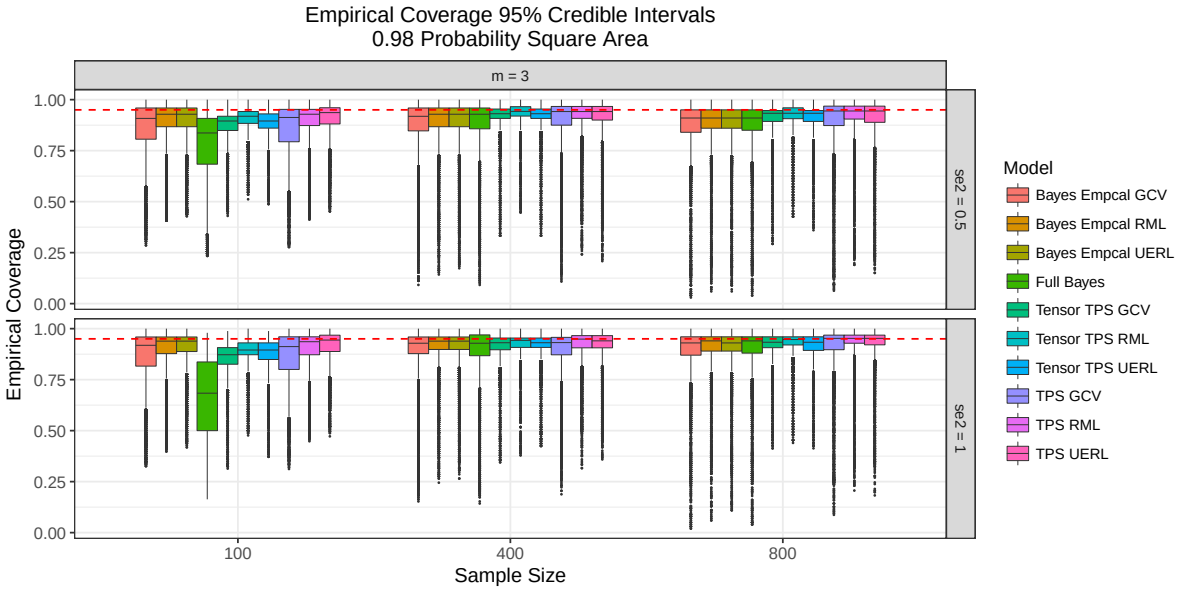

Figure 3.6 Boxplots partial simulation results, empirical coverage of pointwise 95% credible intervals for prediction of multivariate regression functions. Each box is the summary of the empirical coverages $\{\hat{\rho}_i\}_{i=1}^N$. $\hat{\rho}_i \in [0,1]$ is the empirical coverage of the 95% pointwise credible interval for the prediction of $\eta(\chi_i)$ computed after fitting the model to 200 different simulated data sets. The vectors $\{\chi_i\}_{i=1}^N \subset \mathbb{R}^2$ form a grid of resolution 0.05×0.05 in the square $[-2.5, 2.5]^2$. Complete Results in Figure C.5.

If $m < 4$ and the simulated data were generated with $\sigma^2 \leq 0.1$ there is statistical evidence that ζ is different (smaller) than the nominal value for most of the cases. In the cases in which we do not have evidence to reject $\zeta = 0.95$ against $\zeta < 0.95$, we have large variability in the pointwise empirical coverage. Again, the TTPS models provide less variability on the ρ_i's.

The cases $m = 4$ and $\sigma^2 \leq 0.1$ have their one story in terms of empirical coverage, but in practice we would probably not choose such smooth models with $m = 4$ to predict η because these produce the worst predictions and largest variability in the predictions as we discussed from Figures 3.2, 3.3, *C*.1 and *C*.2. Hence in a model selection procedure we would prefer models with $m < 4$.

In general we observe that the ACP, ζ, is close to the nominal value in this simulated example, but we cannot always trust the pointwise empirical coverages to have the nominal levels and, in any of the simulated cases, at most we can expect that half of the pointwise credible intervals have a coverage fairly close to but smaller than the nominal value. Similar results can be observed in plots *C*.6 and *C*.7 where pointwise 60% and 35% credible intervals for the predictions $\eta(\chi_i)$ were computed and the respective coverages were estimated.

We have described the empirical coverage of the pointwise credible intervals for η estimated over the square $[-2.25, 2.25]^2$. As was mentioned before, the square was chosen because it would contain about 98% of the covariates from the simulated data generated using $N_2(\mathbf{0}, I_2)$. Because of this large area of prediction, we have included in our discussions the credible intervals from extrapolation. Next, we reduce the area of estimation of η and compute the empirical coverages for predictions in centered circles around $\mathbf{0} \in \mathbb{R}^2$. We chose the circles of estimation with radius r in a way that they contain about α% of the covariates $\{\mathbf{x}_i\}_{i=1}^n$ generated with the bivariate standard normal distribution. The radius r and α have the relationship $P(z \leq r^2) = \alpha$ with $z \sim \chi^2(2)$ (chi square distribution).

We analyze the distribution of the empirical coverages $\hat{\rho}_i$'s of the credible intervals in these regions as α changes. Plots 3.7 and 3.8 show some of the estimation results. Both sets of boxplots were computed using $m = 3$ in all models and the data generated have a variance error component of $\sigma^2 = 0.5$, 95% credible intervals for the first graphs and 60% credible intervals for the second graphs. More complete results are shown in the appendix Figures C.8 and C.9 which contain the information of the previous Figures and more. Using the same setting as Figure C.9, we show Figure C.10 but now the covariates generated and used for the second set of boxplots were simulated using $\sigma^2 = 0.1^2$.

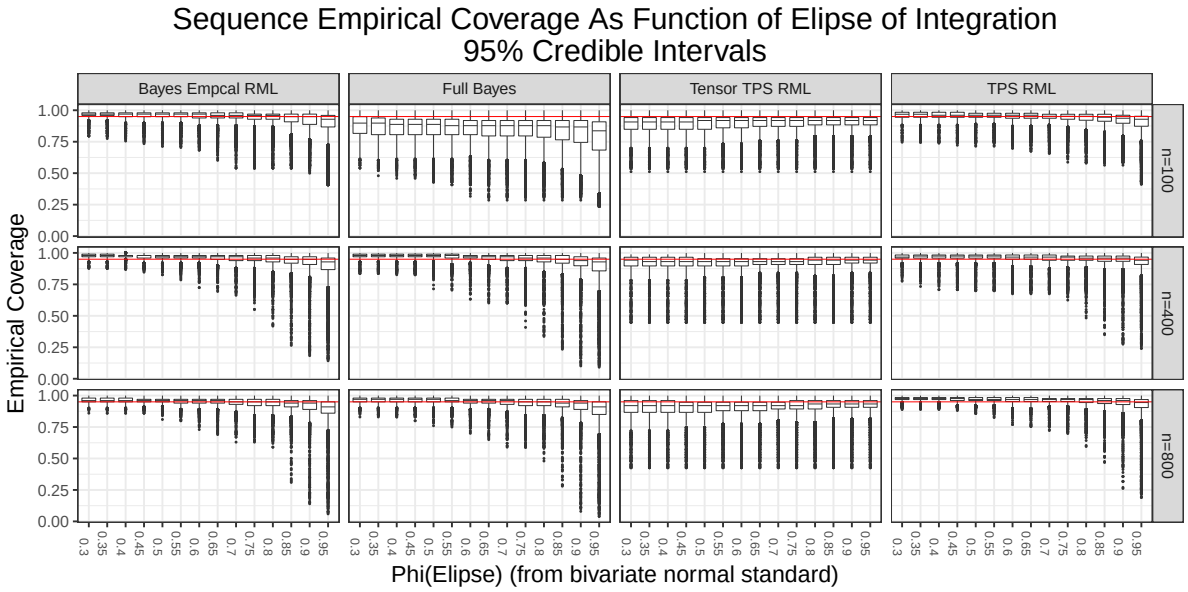

Figure 3.7 Sequential empirical coverage of pointwise 95% credible intervals. *Ellipse* is the ellipse region in $\mathbb{R}^2$ that would contain about $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution. $\alpha = Phi(Ellipse)$. Each box is the the summary of the empirical coverages $\hat{\rho}_i$'s for 95% pointwise credible intervals for the values of $\eta(\chi_i)$ and χ_i in the ellipse region. The sequence is in the sense of observing the distribution of the $\hat{\rho}_i$'s as α changes. The horizontal red line has a value in the vertical axis of 0.95. The simulated data were obtained using $\sigma^2 = 0.5$ and the models were fitted using $m = 3$.

As one can expect when the region of estimation is smaller and is less likely that we are extrapolating (as α decreases, $\alpha = Phi(Ellipse)$ in figures), the certainty of the prediction increases. This property is inherited from the mean of the full conditional posterior of the process η which was constructed with a non-parametric regression method. The method we are using takes advantage of the unbiased estimation of the thin plate splines and tensor thin plate splines which is reflected in the credible intervals being centered in the respective value $\eta(\chi_i)$. But that the empirical coverages $\hat{\rho}_i$'s of the intervals are closer to the nominal value as $Phi(Elipse)$ decreases is a property of the model as a whole. Observe that the extreme whiskers of the boxplots get closer to the nominal value as $Phi(Elipse)$ decreases, especially for the TTPS and TPS models, implying that the pointwise credible intervals for $\eta(\chi_i)$ have the approximate nominal coverage only for small-medium α. From the box plots, α should be smaller of about 0.5 but α should become smaller as n decreases in order to have a fairly

close empirical coverage to the nominal value; it is expected that the non-parametric regression methods estimate incorrectly when n is small. The *Full Bayes* model produces specially under-covering credible intervals when $n = 50$.

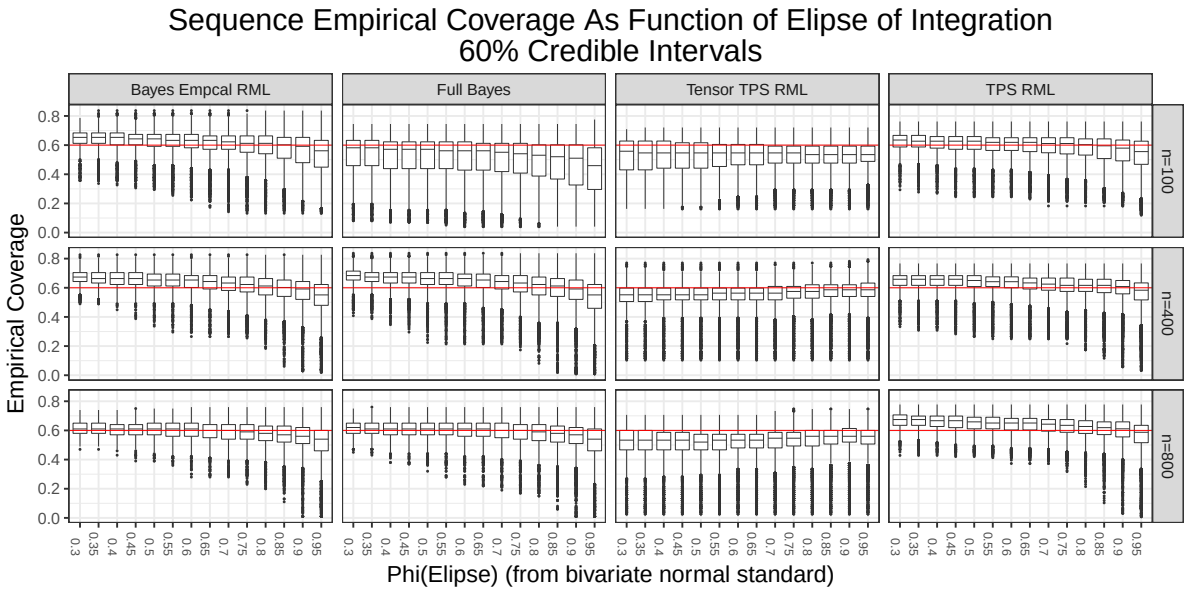

Figure 3.8 Sequential empirical coverage of pointwise 60% credible intervals. *Ellipse* is the ellipse region in $\mathbb{R}^2$ that would contain about $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution. $\alpha = Phi(Ellipse)$. Each box is the the summary of the empirical coverages $\hat{\rho}_i$'s for 95% pointwise credible intervals for the values of $\eta(\chi_i)$ and χ_i in the ellipse region. The sequence is in the sense of observing the distribution of the $\hat{\rho}_i$'s as α changes. The horizontal red line has a value in the vertical axis of 0.60 indicating the nominal coverage. The simulated data were obtained using $\sigma^2 = 0.5$ and the models were fitted using $m = 3$.

We computed and analyzed the empirical coverage of credible intervals with different levels. We only show here and in the appendix the sequence coverages for the 95% and 60% credible intervals but we also computed the same summaries for a finer grid of values for the credibility level C. For the TTPS and TPS models and any method to choose the smoothing parameters we found similar behavior of the empirical coverages regardless of the levels of the intervals when the variance error σ^2 of the response y_i's is not too small ($> 0.1^2$). For $\sigma^2 = 0.1^2$ the intervals from TTPS and TPS models have undercoverage in such degree that not even ζ is close to the nominal value for any value of α, the undercoverage becomes more extreme as σ^2 decreases. We observe a variety of more extreme undercoverage and overcoverage with

the pointwise credible intervals and ζ when using the *Empirical Bayes* models and *Full Bayes* models as σ^2 decreases and for moderate or small sample sizes. The behavior of the credible intervals was hinted at but not evident in Figures 3.6, C.5, C.6 and C.7. The apparent reason for the undercoverage of the credible intervals with TPS and TTPS regardless of the region of estimation (size of α) is that when σ^2 is small we have a fairly good estimation of the function η but the model is overconfident about the estimation and the posterior variance of every $\eta(\chi_i)$ is too small. For now, we did not find a way to modify the variability of the posterior predictive to fix the undercoverage of the credible intervals and preserve the interpretation of the full conditional pointwise estimation as a non parametric regression.

3.4 Conclusions

We reviewed thin plate splines, tensor thin plate splines and linear mixed models to obtain multivariate nonparametric regression methods with normal responses. We reviewed literature for approximation algorithms to faster compute the non parametric regression methods and we used these with four different techniques to choose the bandwidth parameters. We described the frequentist interpretation of all our Bayesian regression methods related to the minimization of a least squared penalized problem; a non-parametric regression method in the frequentist setting. We claim and proved that our estimators of the multivariate regression function, have the same theoretical properties to predict the link functions as the non-parametric regression method. The advantage of our proposed methods is that, besides providing good point estimators, we are able to produce credible intervals for predictions of η. Furthermore we show that our method has the advantage of requiring less storage of the realization of the posterior parameters needed for prediction in relation to other similar Bayesian models.

We set a large simulation study to describe and compare the performance of all the Bayes models to predict the regression function η of two continuous covariates and to study the coverages of the point credible intervals for η evaluated in different regions of the domain. We varied sample sizes and variance errors in the response.

We found that the TTPS and TPS models produce better predictions for the link function using the mean of the posterior predictive distribution and estimate better the variance error

component than the rest of the models. There is no evidence of differences in the performance of predictions within models when using different selections for the bandwidth parameter. We did not find significant difference in the average variability of the predictions around the true function using the variance of the posterior predictive distributions.

The TTPS and TPS models show unbiased estimation of the variance error but only when the sample size is not too small. It is not a surprise that the methods do not perform well for small sample sizes because we are using nonparametric regression methods for the mean of the posterior predictive distribution of η. Poor estimation of the link function induce bad estimation for the variance error parameter. We found as well that for σ^2 to be estimated using any model, it is required that the sample size n increases as σ^2 decreases. In the opposite cases when σ^2 is small and the n is not sufficiently large, the minimum of the realizations of the corresponding posterior distribution were always larger than σ^2. This is equivalent to say, at least from the empirical perspective, that 100% of the empirical credible intervals for σ^2 (not shown in this dissertation) never contained the real value of the variances. It is unfortunate to find this bias in the estimation of the variance produced by all models, but it is a property coming from the rate of covergence of the non parametric methods requiring large sample size.

In the study of the pointwise empirical coverage for the prediction of η we found as well that the average empirical probability ACP is not close to the nominal level when the variance of the errors is too small regardless of the area of prediction. But the TTPS and TPS models seems to produce ACP closer to the nominal value faster than the rest of the models as n increases.

Problem 10 *What is the rate at which the ACP is the nominal value as n increases? does it converge to the nominal value? what happens as* $\sigma^2 \to 0$ *and* $n \to \infty$.

An important feature was observed regarding the coverage of the credible intervals, as these can not be trusted to obtain the nominal level unless the area of prediction is really within the observed covariates (within a radius of 1.2 to the center of the observed covariates in our setting of simulated data, the area corresponds to a region covering about 50% of the data generated from a bivariate standard normal distribution), $n > 100$ and $\sigma^2 > 0.1$. Within this area, about

$80\% - 90\%$ of the credible intervals have an empirical coverage close to the nominal value. One cannot expect the credible intervals in areas of extrapolation to have coverages close to the nominal value but it is necessary to predict within the region of observed data in order to have the desired coverage. Even while all models have these properties, the *Full Bayes* model has specially more deviation from the nominal level to the degree that not even the ACP is close to the nominal value. The *Bayes Empirical* seems to have less such deviations than the *Full Bayes* and is similar to the TTPS and TPS models.

The deviation from the nominal coverage of the credible intervals are in both directions with a tendency of the TTPS and TPS models to undercover. The other two models tend to produce large credible intervals such that they have overcoverage in the case $n \leq 100$ and to produce undercovering credible intervals when $n > 100$.

By the observed results of the simulations, we have provided numerical evidence, at least for the study setting, that the TTPS and TPS models produce better predictions with similar variability on the predictions. The credible intervals for η when using these models preserve the nominal average coverage probability, but the individual intervals do not have the nominal level unless the area of prediction is well within the observed data. We discussed that the TTPS produce better point predictions for η and both methods have similar variability on the predictions. Both models have similar coverages of the credible intervals with the TTPS having a slightly better statistical performance, but as we argued, the difference in the performance of the predictions is not a practical one. We argued that, even when there are statistically significant differences, such differences seem to be of no practical importance: both models produce predictions that detect the general form of the function and the small features of the target function at a similar degree.

The original objective of this chapter was to set the foundations to estimate the link function in a regression problem with errors in the covariates. For the error in the covariates problem, we needed models that can be computed fast and produce good estimation. In Chapter 4, we will theoretically extend the models compared in this chapter to the case with error in the covariates setting, but we choose the TPS model for the task of performing estimation. We found here that the best model to predict in our simulation setting was the TTPS, but this

model has the computational disadvantage of requiring to estimate five smoothing parameters while the TPS model requires only one bandwidth parameter. Given our conclusions that there is no practical difference in the performance of the predictions between these models, we pick the TPS model to extend to the regression problem with measurement errors in the covariates.

CHAPTER 4. BAYESIAN MODEL USING THE APPROXIMATED SOLUTION FOR THE PENALIZED LEAST SQUARES MINIMIZATION PROBLEM IN PRESENCE OF CLASSICAL MEASUREMENT ERROR

A practical example of measurement error in variables arrives in the context of biomarkers. It is known that vitamin D status is associated with bone health, but the mechanism by which vitamin D acts on bones is not completely understood. It is common that researchers work with panel data. For example, panel data of serum bio-marker for vitamin D status, 25-hydroxy D (25(OH)D) and hormone iPTH concentration can be collected for different patients over a long period of time. The iPTH concentration is an indicator of bone health: large serum concentrations of iPTH are associated with poor skeletal health. A question of interest in nutrition epidemiology is how to determine the optimal level of vitamin D status given information on iPTH. By estimating the joint density or the regression function of the biomarkers it is possible to achieve this objective. Due to data collection, additive errors are evident in the measurements of the biomarkers: measurement of the same biomarker on the same person but on different days exhibit large variability caused by, for example the daily and seasonal variability of an individual's diet. Such errors in the measurements cannot be ignored if a reliable density estimation or regression is desired.

Another example of measurement errors is found in relating error-prone predictors such as systolic blood pressure (SBP) to the development of coronary heart disease (CHD). It is known that SBP is measured with error due to the method of collecting the data, and estimates from literature suggest that approximately one third of its observed variability is due to measurement error. In measuring nutrient intake, measurement errors is of concern, impacting on the ability to detect nutritional factors leading to cancer, especially colon and breast cancer. Typical cohort studies measure diet using food frequency questionnaires which, while related

to long term diet, are known to have measurement errors. The failure to account properly for the measurement errors leads to misleading conclusions based on falsely detected statistical significance.

The regression problem with measurement error in the covariates with additive errors in the classical sense is the integration of the usual regression problem, but instead of observing a training set $\{(x_i, y_i)\}_{i=1}^n$, the available data is of the form $\{(w_i, y_i)\}_{i=1}^n$ with $w_i = x_i + \delta_i$, δ_i are the measurement errors with $\mathbb{E}(\delta_i|x_i) = 0$, and x_i and δ_i are independent. The unobserved x is called the *latent variable* and the observed w is called *proxy variable*. The regression function depend on the latent variables and may depend on the proxy covariates as well. Repeated observations of the covariates measured with errors $\left\{\{w_{ij}\}_{j=1}^{n_i}\right\}_{i=1}^n$ may be available and sometimes necessary for identification reasons. For example, if the regression function is assumed to be linear, and the latent variables are not necessarily normal, then the model without repeated proxy observations is identifiable; or if x_i has normal distribution but neither w_i nor δ_i are divisible by a normal distribution, then the model is still identifiable even without repeated observation of the proxy variables, see (Geary (1941); Reiersøl (1950)). In the case of nonlinear regression but known form of the target function and errors in the covariates with vector-valued regressors $\mathbf{x}$, conditions for model identifiability are not known. Schennach et al. (2007) show that in the case of a scalar covariate x, the model is identifiable unless the regression function is of the "log-exponential" form. To the best of our knowledge, there are no general conditions for the regression problem with measurement errors when the regression function is not assumed to have a known form. This setting needs to be studied case by case.

The importance of modeling the measurement errors in the regression problem from theoretical point of view is evident even in the simplest of cases, (Hausman (2001); Fuller (2009)). For example, the simple linear regression model, assuming $Var(x_i) = \sigma_x^2$, when applied naively in the presence of measurement errors with $Var(\delta_i) = \sigma_w^2$, the model produces inconsistent estimation of the slope β as $\hat{\beta} \xrightarrow{n\to\infty} \frac{\beta}{1+\frac{\sigma_w^2}{\sigma_x^2}}$ Fuller (2009). In the limit, the naive least square estimate $\hat{\beta}$ is smaller than the true population slope, an effect called *attenuation*. This effect is illustrated in Figure 4.1.

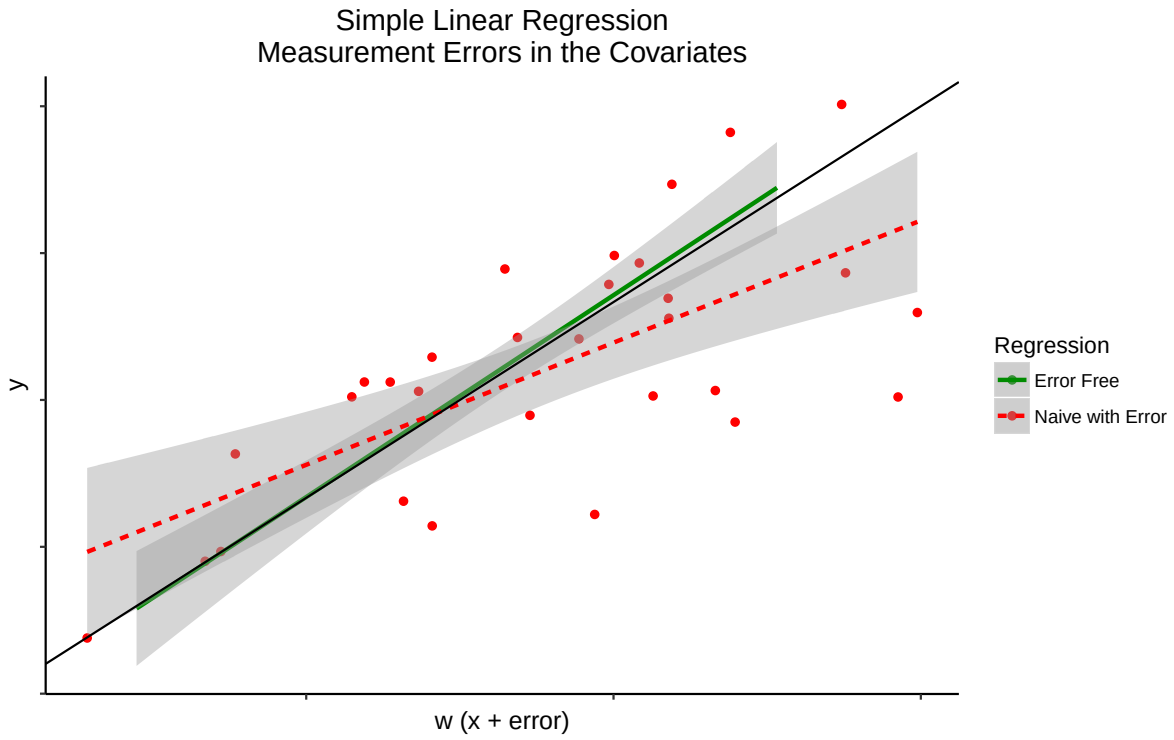

Figure 4.1 Naive linear regression with measurement errors in the covariates. Data simulated using the model $y_i = 1+2x_i+\epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, 1.2^2)$ and $w_i = x_i+\delta_i$ with $\delta_i \overset{iid}{\sim} N(0, 1)$. The pairs $\{(w_i, y_i)\}_{i=1}^{30}$ are shown in the graph. The black line without bands is the true regression function, the regression line *Error Free* is the usual least squares regression with pointwise confidence bands while *Naive with Error* is the fitted regression model using the contaminated values of X, W. Plot computed using software *ggplot2*, function *smooth* Wickham (2009).

There has been, during the last 30 years, a large amount of effort to solve the problem of regression with errors in the covariates. Most of the studies focus on labeled data with standard structure. The settings range from linear regression to nonlinear with known function, nonparametric regression using kernel and deconvolution methods (Stefanski and Carroll (1990)), or methods that estimate the latent covariates such as the EM algorithm or Bayesian approaches. The later two methods use covariates estimates jointly with spline functions. Assumptions on the distributions of the response range from normal distribution with any type of assumption on the regression function to generalized linear regression. With respect to the type of measurement errors, we have found a variety of settings, such as the classical setting (which we have mentioned), the Berkson type measurement error, multiplicative errors, or er-

rors described with known functions. Some examples of the works we found are cited here: Carroll et al. (1984); Stefanski (1985); Stefanski and Carroll (1985); Prentice (1986); Stefanski and Carroll (1987); Schafer (1987); Carroll and Hall (1988); Liu and Taylor (1989); Fan (1990, 1991); Berry et al. (2002); Ruppert et al. (2003); Fuller (2009).

In Sections 4.1, we describe the setting of the regression problem we will consider. In Section 4.2 we describe the models we propose with the proof to all our claims. In order to understand some of the Properties we were not able to proof, we set a simulation study in Section 4.4. Finally, in Section 4.5 we describe regression a model with errors in the covariates and repeated observations in the response.

4.1 Preliminaries

Throughout this chapter, we assume the response variable is univariate normally distributed. We have observations of the continuous covariates $\mathbf{x}$ with errors in the classical sense, $\mathbf{w} = \mathbf{x} + \delta$ with repeated observations $\mathbf{w}$. The regression function may depend on more than one continuous variable, that is to say that $\mathbf{x} \in \mathbb{R}^d$.

Let the training set $\{(\mathbf{w}_{i,1}, \cdots \mathbf{w}_{i,n_i}, y_i)\}_{i=1}^n$ be observed. Without loss of generality we take $n_w := n_i$ for $i = 1, \cdots, n$. We assume the following data generation process

$$\begin{aligned} y_i &= \eta(\mathbf{x}_i) + \epsilon_i \\ \mathbf{w}_{i,j} &= \mathbf{x}_i + \delta_{i,j} \\ \delta_{i,j}|\mathbf{x}_i &\overset{iid}{\sim} N_d(\mathbf{0}, \Sigma_w) \\ \epsilon_i|\mathbf{x}_i &\overset{iid}{\sim} N(0, \sigma^2), \\ \epsilon_i &\perp \mathbf{x}_i, \end{aligned} \tag{4.1}$$

for an unknown function η without necessarily assuming a form, but we assume that $\eta \in \mathcal{H}^\star \subset \mathcal{H}$, with $\mathcal{H}$ a *RKHS* and $\mathcal{H}^\star$ is space of functions of finite low dimension as in Section 2.2. The covariance matrix Σ_w and variance σ^2 are unknown but will be estimated. The assumption of normality on $\delta_{i,j}|\mathbf{x}_i$ and $n_w = 1$ makes the model no identifiable, thus, the condition $n_w > 1$ is required.

The algorithms and methods we propose were strongly inspired by the univariate non parametric *Full Bayes* regression setting in Berry et al. (2002). Berry et al. (2002) proposed three models to estimate a univariate regression function using a Bayes model without assuming a specific form on the regression function. Estimation of the regression function η are achieved using the posterior predictive distribution of η, which conditional to all rest of parameters is a Gaussian process. The problem of choosing the smoothing parameter λ is tackled observing that a linear mixed model can be used as an interpretation of a non parametric regression method. In this way, λ is proportional to the ratio of two variances: the error response variance, and the mixed effect variance. Berry et al. (2002) assigned priors to the variances and the estimation of smoothing parameter follows using standard Bayesian methods. Our Bayesian model from previous chapter is similar to Berry's best model, (out of the three they study), conditional to the unobserved covariates $\{\mathbf{x}\}_{i=1}^{n}$. We extend the model to the multivariate case using thin plate splines. Furthermore, Berry et al. (2002) implicitly use the full form of the solution to a functional minimization problem in a Hilbert space in order to set the mixed effect linear model. We use an approximation to the solution. This approximation is a trade off between extending the model to the multivariate setting and exploring different methods to choose the smoothing parameters. The degree of the approximation is controlled by a unique parameter. which is user defined, so that the approximated solution converges to the exact solution. The form of the solution is fully described by Theorem (52) (the Representer Theorem) as explained in Chapter 2.

Based on the results and discussion of the simulation study presented in Sections 3.3 and (3.4), we do not continue using a full Bayesian multivariate version of Berry model, but instead we use a model related to the approximated solution by means of thin plate splines. Nevertheless, the discussion and results presented here can be analogously extended to the Berry et al. (2002) *Full Bayes* model, to the tensor thin plate splines models previously discussed or to any spline regression.

Most of the theoretical foundations needed for this Chapter and interpretations were already explained in Chapters 2 and 3 with details available in appendix *A*, *B* and *D*. What remains to be done is to state and prove our main theoretical result to fit a non-parametric function

with repeated observations on the covariates measured with errors. A simulation study is set to understand the properties of the proposed model.

As in previous chapters, we are mixing the philosophy of Bayesian statistics and frequentist statistic in our discussions. We are indeed assuming that there is a deterministic function η and deterministic parameters that we try to estimate but we are using Bayesian methods to do so. For example, we use a process defined by a set of distributions $[\eta|\mathbf{y}, \mathbf{W}]$ to estimate the true function η, and the nomenclature is suggestive to indicate that we use this process to estimate η. Similarly for the rest of the parameters in the models.

4.2 Main Theoretical Result

In Proposition 6 we described a Bayes model to estimate the multivariate regression function without measurement errors in the covariates. The interpretation of the fitted regression function from the previous result is that the process defined by the regression's full conditional posterior has a mean that approximately solve the functional minimization problem 1.1 in a previously chosen *RKHS* with its respective induced norm J. We now use this result as a step to generalize the regression model and to deal with the problem in the presence of errors in the covariates. The interpretation of the full conditional posterior process is preserved. We state that the full conditional posterior defines a Gaussian process whose mean function approximates the solution to problem (1.1) conditionally on $\{\mathbf{x}_i\}_{i=1}^n$. We use a process defined by the posterior distribution to estimate the function that relates the latent variables and the response. The rest of the parameters are estimated using the posterior distribution. The proposed model is fully described in Proposition 11.

Proposition 11

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) of functions with domain $\mathbb{R}^d$ and rank in $\mathbb{R}$. Let J the square semi-norm induced by the semi-inner product of $\mathcal{H}$ which has a null space of finite dimension with basis $\{\psi_i\}_{i=1}^l$, let R_J the reproducing kernel of $\mathcal{H}$. Consider $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$. We do not observe $\{\mathbf{x}_i\}_{i=1}^n$ instead, we observe noisy repeated measurements $\{\{\mathbf{w}_{ij}\}_{j=1}^{n_w}\}_{i=1}^n$ (with out lost of generality n_w does not depend on i) with $\mathbf{w}_{ij} = \mathbf{x}_i + \delta_{ij}$

and $\delta_{ij}|\mathbf{x}_i \overset{iid}{\sim} N_d(\mathbf{0}, \Sigma_w)$. *The labels* i *and* j *are known. Let* $\mathbf{Z} := \{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n =: \mathbf{X}$, $\mathbf{d} := (d_1\, d_2\, \cdots\, d_l)^\intercal \in \mathcal{M}_{l\times 1}(\mathbb{R})$, $\mathbf{c} := (c_1\, c_2\, \cdots\, c_k)^\intercal \in \mathcal{M}_{k\times 1}(\mathbb{R})$. *Consider the model*

$$\begin{aligned} y_i &= \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i) + \epsilon_i, \\ \eta_{\binom{\mathbf{d}}{\mathbf{c}}} &= \sum_{i=1}^{l} d_i\psi_i + \sum_{i=1}^{k} c_i R_J\left(\mathbf{z}_i, \cdot\right) \\ \epsilon_i|\mathbf{x}_i &\overset{iid}{\sim} N_1(0, \sigma^2). \end{aligned}$$

Let $\lambda|\mathbf{X} > 0$, *and* $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ *with entries* $Q_{i,j} = R_J(\mathbf{z}_i, \mathbf{z}_j)$, $S|\mathbf{X} \in \mathcal{M}_{n\times l}(\mathbb{R})$ *with* $S_{i,j}|\mathbf{X} = \psi_j(\mathbf{x}_i)$ *full column rank,* $R|\mathbf{X} \in \mathcal{M}_{n\times k}(\mathbb{R})$, $R_{i,j}|\mathbf{X} = R_J(\mathbf{x}_i, \mathbf{z}_j)$, $M|\mathbf{X} = RQ^+R^\intercal + n\lambda I_n$, $m_x > 0$. *Consider the priors*

$$\begin{aligned} d_i &\overset{iid}{\sim} 1, \\ \mathbf{c}|\sigma^2, \lambda &\sim N_k\left(\mathbf{0}, \frac{\sigma^2}{n\lambda}Q^+\right), \\ \mathbf{P}(\lambda \geq \lambda_0|\mathbf{X} = \mathbf{x}, \sigma^2 = s^2) &= \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{U}(x|\mathbf{y}, \mathbf{X} = \mathbf{x}, \sigma^2 = s^2)\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}), \end{aligned} \tag{4.2}$$

$$\sigma^2 \sim Inv - Gamma(A_\epsilon, B_\epsilon),$$

$$\begin{aligned} \mathbf{x}_i|\boldsymbol{\mu}_x, \Sigma_x &\overset{iid}{\sim} N_d(\boldsymbol{\mu}_x, \Sigma_x),\ i = 1, .., n, \\ \boldsymbol{\mu}_x|\Sigma_x &\sim N_d(\mathbf{d}_x, m_x^{-1}\Sigma_x), \\ \Sigma_x &\sim Inv - Wishart(\mathbf{A}_x, b_x), \\ \Sigma_w &\sim Inv - Wishart(\mathbf{A}_w, b_w), \\ \mathbf{d}|\mathbf{X} \perp \mathbf{c}|\mathbf{X}, \quad \mathbf{d}|\mathbf{X} \perp \sigma^2|\mathbf{X}, \quad &\binom{\mathbf{d}}{\mathbf{c}}|\mathbf{X} \perp (\epsilon_1 \cdots \epsilon_n)^\intercal\,|\mathbf{X}. \end{aligned}$$

Unless stated otherwise the rest of the parameters have independent priors. Alternatively, we could assign either of the following conditional priors to λ*:*

$$\mathbf{P}(\lambda \geq \lambda_0|\mathbf{X} = \mathbf{x}) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{V}(x|\mathbf{y}, \mathbf{X} = \mathbf{x}, \alpha = 1.4)\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}),\ \textit{or} \tag{4.3}$$

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X} = \mathbf{x}) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{M}(x|\mathbf{y}, \mathbf{X} = \mathbf{x})\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}). \tag{4.4}$$

Then the joint posterior of the parameters exists and the full conditional posteriors are

- $\binom{\mathbf{d}}{\mathbf{c}}|\mathbf{y},\ \sigma^2,\ \lambda,\ \mathbf{X} \sim N_{l+k}\left(\mu_{\mathbf{dc}}, \frac{\sigma^2}{n\lambda}\mathbf{\Sigma_{dc}}\right)$, *where*

$$\mu_{\mathbf{dc}} = \begin{pmatrix} (S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1} \\ Q^+ R^\intercal M^{-1}(I - S(S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1}) \end{pmatrix} \mathbf{y}$$

$$\mathbf{\Sigma_{dc}} = \begin{pmatrix} (S^\intercal M^{-1} S)^{-1} & -(S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1} R Q^+ \\ -Q^+ R^\intercal M^{-1} S (S^\intercal M^{-1} S)^{-1} & Q^+ - Q^+ R^\intercal \{M^{-1} - M^{-1} S (S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1}\} R Q^+ \end{pmatrix}$$

- $\sigma^2|\mathbf{y},\ \binom{\mathbf{d}}{\mathbf{c}},\ \mathbf{X} \sim Inv - Gamma\left(A_\epsilon + \frac{1}{2}n,\ \left[B_\epsilon^{-1} + \frac{1}{2}\sum_{i=1}^n \left(y_i - \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i)\right)^2\right]^{-1}\right)$,

-

$$\begin{aligned} \mathbf{x}_i|\mathbf{y},\ \binom{\mathbf{d}}{\mathbf{c}} &\propto [\mathbf{y}_i|, \sigma^2, \mathbf{x}_i] \prod_{j=1}^{n_w} [\mathbf{w}_{ij}|\mathbf{x}_i, \Sigma_w]\, [\mathbf{x}_i|\boldsymbol{\mu}_x, \Sigma_x] \\ &= \exp\left\{-\frac{1}{2\sigma^2}\left\|\mathbf{y}_i - \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i)\right\|^2 - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_x)'\Sigma_x^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_x)\right\} \\ &\quad \times \exp\left\{-\frac{1}{2}\sum_{j=1}^{n_w}(\mathbf{w}_{ij} - \mathbf{x}_i)'\Sigma_w^{-1}(\mathbf{w}_{ij} - \mathbf{x}_i)\right\}, \end{aligned}$$

- $\Sigma_w|\mathbf{y}, \mathbf{x}_1, \cdots, \mathbf{x}_n, \Sigma_x \sim Inv - Wishart\left[\mathbf{A}_w + \boldsymbol{A}\boldsymbol{A}', nn_w + b_w\right]$
  where $\mathbf{A} = [\mathbf{x}_1 - \mathbf{w}_{11}..\mathbf{x}_1 - \mathbf{w}_{1n_w} \cdots \mathbf{x}_n - \mathbf{w}_{n1}..\mathbf{x}_n - \mathbf{w}_{n,n_w}]$,

- $\boldsymbol{\mu}_x|\mathbf{y}, \Sigma_x, \mathbf{x}_1, \cdots, \mathbf{x}_n \sim N_d\left[\Sigma_x^{-1}\left(n\bar{\mathbf{x}} + m_x\mathbf{d}_x\right), \frac{1}{n+m_x}\Sigma_x\right]$, *with* $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i$,

- $\Sigma_x|\mathbf{y}, \mathbf{x}_1, \cdots, \mathbf{x}_n = Inv - Wishart\left[\mathbf{A}_x + n\mathbf{S} + \frac{nm_x}{n+m_x}(\bar{\mathbf{x}} - d_x)(\bar{\mathbf{x}} - d_x)', n + b_x\right]$
  where $\mathbf{S} = n^{-1}\sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$,

- $(\lambda|\mathbf{y} = y, \mathbf{X} = \mathbf{x}, \sigma^2 = s^2) = \arg\min_{x>0} \mathcal{U}(x|\mathbf{y} = y, \mathbf{X} = \mathbf{x}, \sigma^2 = s^2)$ *almost surely.*

If the other priors on λ *depending on* $\mathcal{V}$ *or* $\mathcal{M}$, (4.3) *or* (4.4) *were assigned, the corresponding full conditional posterior distributions are:*

$$(\lambda|\mathbf{y} = y, \mathbf{X} = \mathbf{x}) = \arg\min_{x>0} \mathcal{V}(x|\alpha = \alpha_0, \mathbf{X} = \mathbf{x}, \mathbf{y} = y) \text{ almost surely, or} \tag{4.5}$$

$$(\lambda|\mathbf{y} = y, \mathbf{X} = \mathbf{x}) = \arg\min_{x>0} \mathcal{M}(x|, \mathbf{X} = \mathbf{x}, \mathbf{y} = y) \text{ almost surely.} \tag{4.6}$$

Proof.

The existence of the posterior would be direct if all the priors were proper but since we decided to assign an improper prior to the vector $\mathbf{d}$ this is not the case. The proof is tedious but direct. The rigorous steps follow the lines from Proposition 6 and Section 3.2.1 which uses Propositions 72 and 74. Lets denote $\boldsymbol{\Theta} = \left(\mathbf{d}^\intercal,\, \mathbf{c}^\intercal,\, \sigma^2\,\Sigma_w,\, \mu_x, \Sigma_w, \lambda\right)^\intercal$. By the same arguments as in Proposition 6, it can be shown that the posterior exists and it is written as

$$[\boldsymbol{\Theta}|\mathbf{y},\, \mathbf{W}] \propto \left[\mathbf{y}|\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right), \mathbf{X}, \sigma^2, \right] \times [\mathbf{W}|\Sigma_w\, \mathbf{X}] \times \left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)|\sigma^2,\, \mathbf{X}, \lambda\right] \times \left[\lambda|\sigma^2, \mathbf{X}\right]$$
$$\times\, [\mathbf{X}|\mu_x, \Sigma_x] \times [\mu_x|\Sigma_x]\, [\Sigma_w]\, [\Sigma_x] \times \left[\sigma^2\right].$$

By writing explicitly the posterior distribution, reordering terms, using the conjugacy property of the inverse Wishart distribution with respect to the multivariate normal distribution, completing terms, we found the full conditional posteriors distributions. The process is tedious but the non trivial arguments, conditional to $\mathbf{X}$, were derived in Section 3.2.1. ■

One of the main focus of this dissertation is the estimation of the regression function η, but the rest of the parameters need to be estimated as well. Observe that the full conditional posterior of $\mathbf{d}$ and $\mathbf{c}$ in Proposition 11, which are used directly to estimate η, depends on σ^2. A good estimate of σ^2 is thus required. We decided to use a prior inverse-gamma $[\sigma^2]$ because of the conjugacy of the distribution. This prior leads to a full conditional posterior on σ^2 that depends on an "average" of the differences of the observed y_i's and the current estimate of η evaluated at the current estimates of $\{\mathbf{x}_i\}_{i=1}^n$ of η, and consequently computing a "mean squared error" expression. On the other side, the smoothing parameter λ is conditionally chosen to the current estimates of the latent variables, by minimizing the score functions $\mathcal{U}$, $\mathcal{V}$ or $\mathcal{M}$ that approximate a weighted sum of the mean squared error and a penalty term. These considerations lead to evidence that the full conditional estimate of σ^2 may be biased in the sense that the mean of the marginal posterior $[\sigma^2|\mathbf{y}, \mathbf{W}]$ is not the true parameter σ^2. In an attempt to improve the overall estimation by a priori improving the estimation of σ^2, we can use methods which main focus is to estimate is to estimate this parameter in the classical regression problem. One type of frequentist methods is to provide estimators for the variance by providing an estimator for the regression function η, see (Spokoiny (2002); Tong

et al. (2013); Zhou et al. (2015)); while another type of estimators called difference methods, do not estimate η (Devroye et al. (2003); Liitiäinen et al. (2008, 2009, 2010); De Brabanter et al. (2014)). There are others papers that estimate the variance component as well in the presence of measurement errors, (Delaigle and Hall (2011); Delaigle (2014)), but these methods use deconvolution and kernel algorithms which would be difficult to incorporate them to our current Bayesian setting. The estimator of the variance in the classical linear regression setting with errors in the covariates is addressed in Fuller (2009). These estimators for the observation-error variance σ^2 in the regression problem without measurement error in the covariate, can be incorporated in a model similar to the one described in Proposition 11. For example, using k-th nearest neighbors (Definition 44) instead of the inverse gamma prior-full conditional distributions, we would like to have either of the following full conditional posteriors $[\sigma^2|\mathbf{X},\mathbf{y}]$

$$\left(\sigma^2|\mathbf{X}=\mathbf{x},\mathbf{y}=(y_1,\cdots,y_n)\right)=\frac{1}{n}\left(\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}y_iy_{i,1}\right)\ \text{almost surely, or} \tag{4.7}$$

$$\left(\sigma^2|\mathbf{X}=\mathbf{x},\mathbf{y}=(y_1,\cdots,y_n)\right)=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-y_{i,1}\right)^2\ \text{almost surely, or} \tag{4.8}$$

$$\left(\sigma^2|\mathbf{X}=\mathbf{x},\mathbf{y}=(y_1,\cdots,y_n)\right)=\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-y_{i,1}\right)\left(y_i-y_{i,2}\right)\ \text{almost surely, or} \tag{4.9}$$

$$\left(\sigma^2|\mathbf{X}=\mathbf{x},\mathbf{y}=(y_1,\cdots,y_n)\right)=\frac{1}{nk^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{k}\left(y_i-y_{i,2,j}\right)\right)\left(\sum_{j=1}^{k}\left(y_i-y_{i,2,j-1}\right)\right)\ \text{a.s., } k>1, \tag{4.10}$$

where the notation is conserved from Proposition 11. In this way, we preserve the notion that, given the latent variables $\mathbf{X}$ we can estimate the true σ^2 using the expressions above. For the incorporation of these ideas to a Bayesian model, it is needed to show the existence of priors that lead to the conditional distributions above. The priors can be implicitly obtained as follow

$$\mathbf{P}\left(\sigma^2\geq s^2|\mathbf{X}=\mathbf{x}\right)=\int_{\mathbb{R}^n}\mathbf{1}\left\{s^2\geq\frac{1}{n}\left(\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}y_iy_{i,1}\right)\right\}dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}),\ \text{or} \tag{4.11}$$

$$\mathbf{P}\left(\sigma^2\geq s^2|\mathbf{X}=\mathbf{x}\right)=\int_{\mathbb{R}^n}\mathbf{1}\left\{s^2\geq\frac{1}{n}\sum_{i=1}^{n}\left(y_i-y_{i,1}\right)^2\right\}dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}),\ \text{or} \tag{4.12}$$

$$\mathbf{P}\left(\sigma^2 \geq s^2 | \mathbf{X} = \mathbf{x}\right) = \int_{\mathbb{R}^n} \mathbf{1}\left\{s^2 \geq \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - y_{i,1}\right)\left(y_i - y_{i,2}\right)\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}), \text{ or} \tag{4.13}$$

$$\mathbf{P}\left(\sigma^2 \geq s^2 | \mathbf{X} = \mathbf{x}\right) = \int_{\mathbb{R}^n} \mathbf{1}\left\{s^2 \geq \frac{1}{nk^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{k}\left(y_i - y_{i,2,j}\right)\right)\left(\sum_{j=1}^{k}\left(y_i - y_{i,2,j-1}\right)\right)\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}),\ k > 1, \tag{4.14}$$

where $F_{\mathbf{y}|\mathbf{X}=\mathbf{x}}$ is the upper tail of the conditional cumulative distribution of $\mathbf{y}$ given $\mathbf{X} = \mathbf{x}$.

Other options available for full conditional posteriors $[\sigma^2|\mathbf{X}, \lambda, \mathbf{y}]$ for the variance component can be proposed as well using the estimators described by (2.60) and (2.73) as:

$$\left(\sigma^2 | \mathbf{X} = \mathbf{x}, \lambda = \lambda, \mathbf{y} = \mathbf{y}\right) = \sigma_v^2(\lambda, \mathbf{x}, \mathbf{y}) \text{ almost surely, or} \tag{4.15}$$

$$\left(\sigma^2 | \mathbf{X} = \mathbf{x}, \lambda = \lambda, \mathbf{y} = \mathbf{y}\right) = \sigma_m^2(\lambda, \mathbf{x}, \mathbf{y}) \text{ almost surely,} \tag{4.16}$$

with corresponding priors:

$$\mathbf{P}\left(\sigma^2 \geq s^2 | \mathbf{X} = \mathbf{x}, \lambda = \lambda\right) = \int_{\mathbb{R}^n} \mathbf{1}\left\{s^2 \geq \sigma_v^2(\lambda, \mathbf{x}, \mathbf{y})\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}, \lambda=\lambda}(\mathbf{y}) \tag{4.17}$$

$$\mathbf{P}\left(\sigma^2 \geq s^2 | \mathbf{X} = \mathbf{x}, \lambda = \lambda\right) = \int_{\mathbb{R}^n} \mathbf{1}\left\{s^2 \geq \sigma_m^2(\lambda, \mathbf{x}, \mathbf{y})\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}, \lambda=\lambda}(\mathbf{y}). \tag{4.18}$$

Thus, six more Bayesian models similar to the one described in Proposition 11 with the same full conditional distribution $[\eta(\chi)|\mathbf{y}, \lambda, \mathbf{d}\,\mathbf{c}\,\sigma^2]$ can be proposed by changing the inverse-gamma prior $[\sigma^2]$ by either of the six priors (4.11) – (4.18). The existence of the joint posterior distribution of all parameters in such models would need to be proven to exists because $\mathbf{d}$ has still an improper prior $[\mathbf{d}] \sim 1$. Existence of the joint posterior distribution would follow using the same steps as in Proposition 11. The full conditional posterior distribution of the parameters in the six Bayesian models are same with the exception of $[\sigma^2|\mathbf{y}, \mathbf{X}, \lambda]$; such distributions for $\sigma^2|\mathbf{y}, \mathbf{X}, \lambda$ are described by (4.7) – (4.15) respectively.

For observed latent variables $\mathbf{X}$ in the frequentist setting, the difference variance estimator $\frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i y_{i,1}\right)$ used in (4.7) was proposed by De Brabanter et al. (2014), the corresponding frequentist estimator used in (4.8) was proposed by Devroye et al. (2003), and the estimators used in (4.9) and (4.10) were proposed by Liitiäinen et al. (2009, 2010). Conditionally on the latent variables, it was proven by the corresponding authors that under some

regularity conditions (conditions hold in our case for ϵ_i normal distributed), the rate of convergence to zero of $\mathbb{E}\left|\hat{\sigma}_n^2 - \mathbb{E}e^2\right|$ is at most of order $c_1 n^{-1/2} + c_2 n^{-2/d}$. Such rate of convergence is indeed lower than the rate convergence of estimators of variances using estimators of the regression function. Still, we hypothesized that using these priors for σ^2 in the context to the errors invariables problem we may have an improvement because we only require to use the estimator of the latent variables in comparison with the inverse gamma prior that needs as well estimator of the regression function η. We test the hypothesis using simulations in the following sections.

In regard to the prior proposed for the latent variable $\mathbf{x}_i$'s, we think that there is not a reasonable distribution in the general case. The prior may be changed according to the required application. A flat reference prior may be reasonable; normal hierarchical could be another option, mixture of normal distributions is a flexible prior but in this case we may have problems identifying the group of the mixture to where an observation belongs, Gelman et al. (2014), and the updates in the MCMC algorithm may be slow. The prior distribution is indeed an open choice for each particular problem. In what follows, we use a hierarchical normal prior.

4.3 Implementation and Interpretation

Estimators for all parameters in the model described in Proposition 11; or the models with the variations on the priors (4.11), (4.12), (4.13) or (4.14) are obtained using the posterior distribution while any predictions on the value η at any set of points $\{\chi\}_{i=1}^N \subset \mathbb{R}^d$ are obtained using the posterior predictive distribution. (Gelman et al., 2014, p. 145):

$$\Pi(\eta(\chi)|\mathbf{y}, \mathbf{W}) = \int \Pi\left(\eta(\chi)|\theta\right) \Pi\left(\theta|\mathbf{Y}\right) d\theta.$$

Again, by simplicity of notation as in Section 3.3 we will abuse of the notation: η is a deterministic function and we use the process $\Pi(\eta|\mathbf{y}.\mathbf{W}) =: [\eta|\mathbf{y}, \mathbf{W}]$ to estimate it. We consider as point estimator for $\eta(\chi)$ to be the mean of the posterior predictive distribution $[\eta(\chi)|\mathbf{y}, \mathbf{W}]$.

Samples from the posterior distribution of the parameters are drawn using a MCMC method and realizations $\{\hat{\eta}(\chi)_i\}_{i=1}^M$ from $\Pi(\eta(\chi)|\mathbf{y}, \mathbf{W})$ are obtained using the draws $\left\{ \begin{smallmatrix} \hat{\mathbf{d}}_i \\ \hat{\mathbf{c}}_i \end{smallmatrix} \right\}_{i=1}^M$ of the

corresponding posteriors followed by computing

$$\hat{\eta}(\chi)_i = \sum_{j=1}^{l} d_j^{(i)} \psi_j(\chi) + \sum_{j=1}^{k} c_j^{(i)} R_J\left(\mathbf{z}_j, \chi\right).$$

The known functions $\{\psi_i\}_{i=1}^{l}$ are a basis for the null space of the semi norm J and R_J is the Reproducing Kernel of a *RKHS* as specified in Proposition 11. For an exposition on *RKHS* please revisit Chapter 2. Computation of R_J for the thin plate splines can be accomplished for example, by using the description provided in Section 2.1.1.2. For the computation of R_J in case of a tensor smoothing with thin plate splines as marginals can be accomplished by building on the tensor thin plate splines case as described in Section 2.1.2. The size k of the sets of knots $\{\mathbf{z}_i\}_{i=1}^{k}$ is chosen to be

$$k = \lfloor \max\left\{30, 10n^{2/9}\right\} \rfloor$$

as described in Section 2.2.3.

Proposition 11 requires that $\{\mathbf{z}_i\}_{i=1}^{k} \subset \{\mathbf{x}_i\}_{i=1}^{n}$ but the latent variables are not observed. The knots are used to define the space $\mathcal{H}^\star \subset \mathcal{H}$ where the mean of the full conditional process belongs to. We use the averages $\{n_w^{-1} \sum_{j=1}^{n_w} \mathbf{w}ij\}_{i=1}^{n}$ as if they were the true latent variables and the knots are chosen as described in Section 2.2.3. The averages are only used to implement the method to choose the knots. We run some simulated examples and for these cases it seems that using this method still produces similar results as if we had a subset of the latent variables. We recognize that more studies need to be done to theoretically or numerically evaluate the veracity of this method.

The posterior distribution of the parameters in the models and the posterior predictive distribution $\Pi(\eta(\chi)|\mathbf{Y})$ do not have an analytical form; at least, we did not identify and explicit form. Samples from the k=joint posterior distribution were drawn using the MCMC method Metropolis Hasting within Gibbs sampler (Gelman et al. (2014)). Observations from each of the full conditional distributions are drawn in the order they appear in Proposition 11. The generation of an observation $\left[\binom{\mathbf{d}}{\mathbf{c}} | \mathbf{y}, \sigma^2, \lambda, \mathbf{X}\right]$ is computationally expensive because the matrices R and S depend on $\{\mathbf{x}_i\}_{i=1}^{n}$ and on λ which are continually changing in the algorithm. A realization of $\left[\lambda|\mathbf{X}, \sigma^2\right]$ is computationally expensive as well because the scores functions $\mathcal{M}$, $\mathcal{V}$ and $\mathcal{U}$ are continually changing as well in each iteration of the MCMC. In fact, we do not use

the definition of these scores to draw a λ value because the evaluation of the scores are slow and computationally unstable. We refer to appendix *D* for a more comprehensive explanation on this topic.

The generation from the full conditional posterior of $\mathbf{x}_j$ requires a Metropolis-Hasting step. We use an *adaptive Metropolis* algorithm proposed by Roberts and Rosenthal (2009). At the i^{th} iteration, the proposal distribution for $i \leq 2d$ is

$$Q_i(\mathbf{x}) = N(\mathbf{x}, 0.1^2 I_d c_i/d);$$

and for $i > 2d$ we use

$$Q_i(\mathbf{x}) = \begin{cases} (1-\theta)N(\mathbf{x}, 2.38^2\tilde{\Sigma}_i c_i/d) + \theta N(\mathbf{x}, 0.1^2 I_d c_i/d) & \tilde{\Sigma}_i > 0 \\ N(\mathbf{x}, 0.1^2 I_d c_i/d) & \tilde{\Sigma}_i \not> 0 \end{cases}$$

for some $\theta \in (0,1)$ and the empirical covariance matrix

$$\tilde{\Sigma}_i = \frac{1}{i}\left(\sum_{j=0}^{i} \mathbf{x}_j \mathbf{x}_j^\intercal - (i+1)\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\intercal\right),$$

where $\bar{\mathbf{x}}_i = \frac{1}{i+1}\sum_{j=0}^{i} \mathbf{x}_j$, $c_i = \exp\left(\min(1^{-2}, n^{-\frac{1}{2}})(\mathbf{1}_{accpt_i > .44} - \mathbf{1}_{accpt_i < .44})\right)$. We use $\theta = 0.1$. According to Gelman et al. (2014), the acceptance-rejection proportion of the proposed new value in the Metropolis step is desired to be just below 0.44 for the dimension of $\mathbf{x}_i$ being $d = 2$ but as $d > 4$ the proportion desired is about 0.23. Figure 4.2 show the proportion of acceptance for each $\mathbf{x}_i$, $i = 1, \ldots, 100$ after the burn-in period that we obtained in a usual run. The generation of the parameters, besides the coefficients $\mathbf{d}$, $\mathbf{c}$, the smoothing parameter λ and the Metropolis part, are computationally straight forward. We save the realizations of all these parameters.

Two independent chains with different initial overdispersed values for each parameter were drawn using Metropolis-Hasting within Gibbs sampler as described. Each of the chains were run for $15,000$ iterations discarding the first $10,000$ realizations as burn-in and thinning the rest of the sequences by keeping every 3 draws. Convergence tests such as Geweke test, (Geweke et al. (1991)) and Gelman test (Gelman et al., 2014, p. 285) were used to separately test the convergence of the chains of each parameter.

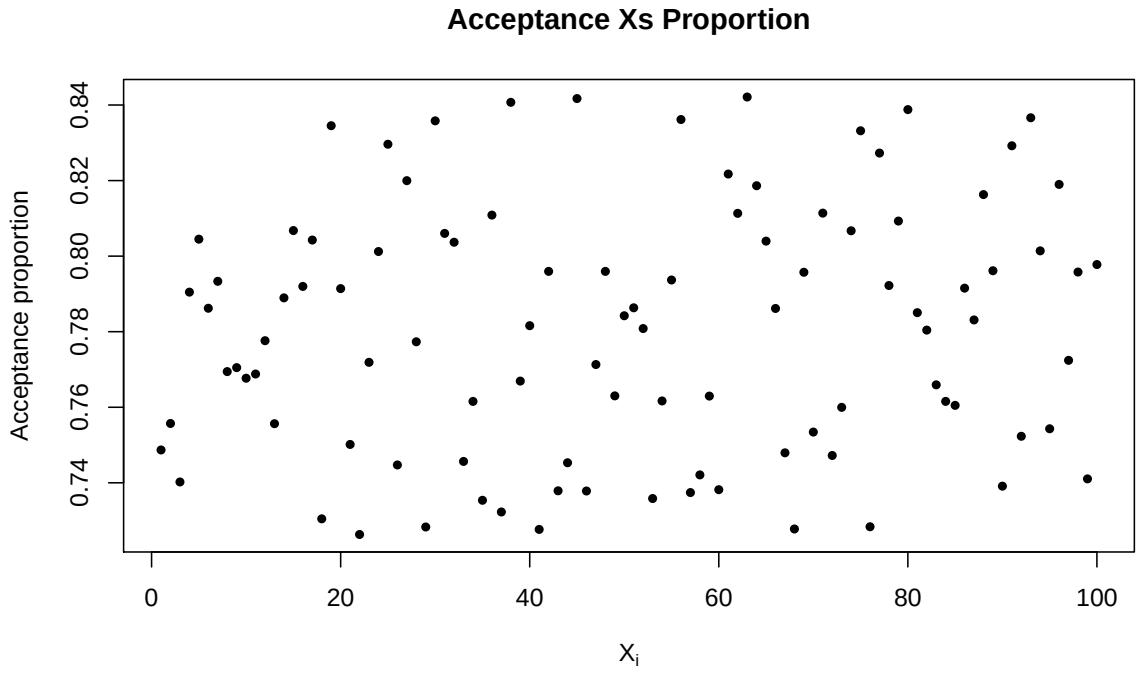

Figure 4.2 Acceptance rate of the Metroplis-Hastings step after the burn in period for an example of estimation. In this example $n = 100$ latent variables are estimated. The acceptance rate is larger than the desired.

For $\chi \in \mathbb{R}^d$, let $\hat{\eta}(\chi) := \mathbb{E}\left[\Pi(\eta(\chi)|\mathbf{Y})\right]$ the point estimates of $\eta(\chi)$ and $\tilde{\eta}(\chi) := \mathrm{sd}\left[\Pi(\eta(\chi)|\mathbf{Y})\right]$ the standard deviation of the posterior predictive distribution. $\hat{\eta}(\chi)$ and $\tilde{\eta}(\chi)$ are estimated with the sample mean and the unbiased sample standard deviation from the realizations of the posterior predictive distributions.

Figure 4.3 shows two examples of the regression estimation using the posterior predictive of $\eta(\chi_i)$ from the model in Proposition 11. The used reproducing kernel R_J is the corresponding of the thin plate splines (Section 2.1.1.2) problem. The plotted predictions $\hat{\eta}(\chi_i)$ are in a grid of resolution 0.05×0.05 inside the square $[-2.25, 2.25]^2$. The grid is denoted as $\{\chi_i\}_{i=1}^N \subset \mathbb{R}^2$. The square $[-2.25, 2.25]^2$ was chosen to be used as the region of estimation for all the simulated data sets of the study because it has the property that it would contain about 95.17% of all points simulated from $N_2(\mathbf{0}, I_2)$; the grid was not chosen to be finer because of storage availability. The plot at the left-top contains the level curves of the true regression function, the center column of plots are the corresponding prediction and variability of prediction obtained by fitting a model

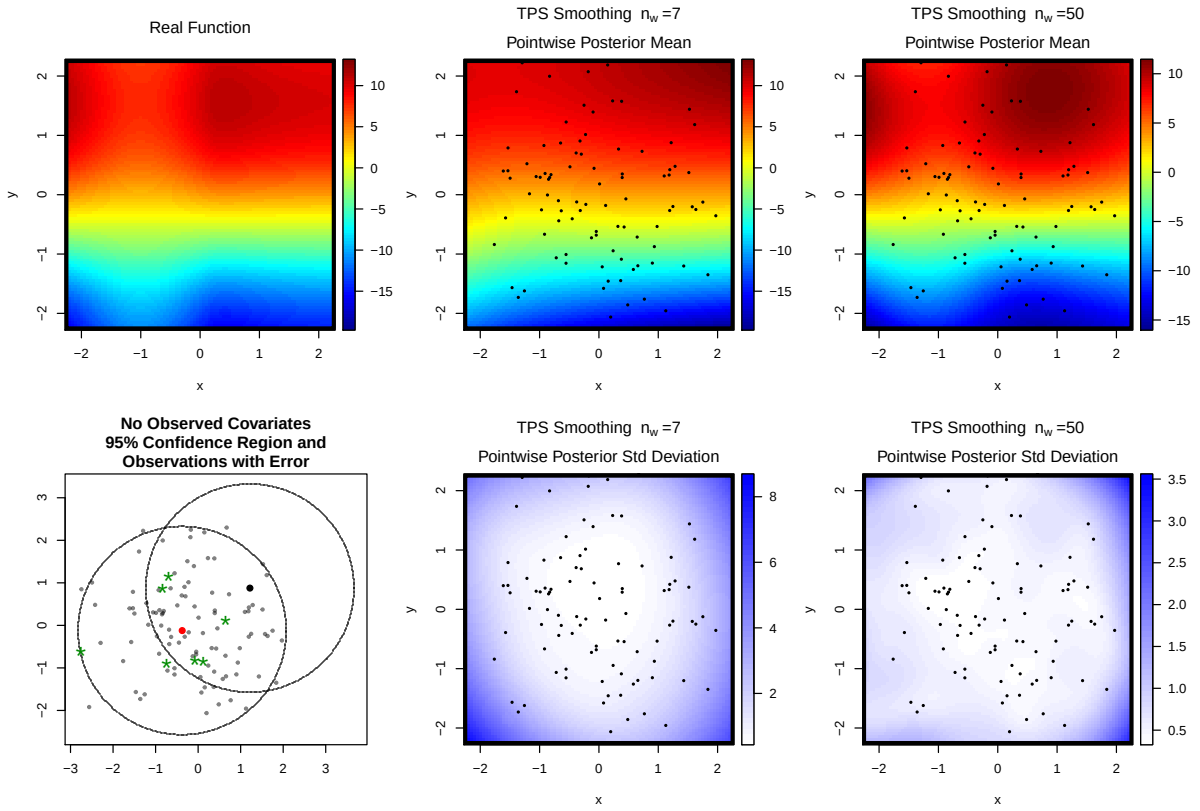

Figure 4.3 Example estimate and standard deviation of the estimator for the multivariate regression problem with measurement error in the covariates. Level curves for the true function η (3.17) (top left plot). Point Bayes estimate $\hat{\eta}(\chi)$ (top center and right), and pointwise standard deviation, $\tilde{\eta}(\chi)$ (bottom center and right). Estimation using a Bayesian model interpretation of the thin plate splines with $m = 3$ and smoothing parameter chosen using the restricted maximum likelihood method. Data simulated with $n = 100$, $\sigma^2 = 0.5$, $\Sigma_w = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. The dots in the plots are the unobserved values of the covariates $\{\mathbf{x}_i\}_{i=1}^{100}$. The plot in the bottom left shows the covariates $\{\mathbf{x}_i\}_{i=1}^{100}$ and for two $\mathbf{x}_i$ and $\mathbf{x}_j$, the 95% confidence regions are shown where the corresponding $\mathbf{w}_{ij}$ are expected to be observed. Inside one of the regions, a set of $\{\mathbf{w}_{ij}\}_{j=1}^{7}$ are plotted.

to data with $n_w = 7$ repeated measures with error of the latent variables; the right columns, for $n_w = 50$ repeated measures with errors of the latent variables. The top center and right plots are the predictions $\hat{\eta}$, and the bottom center and right plots are the level curves for the standard deviation $\tilde{\eta}$. The dots in plots are the latent variables, they are not used for the estimation. The pointwise standard deviation of the estimation $\tilde{\eta}$ is larger on the boundaries of the region of estimation because we do not have information of the function in that area; while the standard deviation is smaller in the center of the region as expected. The graph at the bottom left of Figure 4.3 shows the latent variables (gray dots), the circles centered in two realizations $\mathbf{x}_i$ and $\mathbf{x}_j$ would contain about 95% of the observed measurements with error $\{\mathbf{w}_{i,\nu}\}_{\nu=1}^{n_w}$. The green dots are the measurements with error of the latent variable represented with a red dot $\mathbf{x}_i$.

4.4 Simulation Study

We design a simulation study to compare the performance of the models in Section 4.2 in terms of point estimates and coverage of credible intervals for functions $\eta : \mathbb{R}^2 \to \mathbb{R}$. As in the free error case, the algorithm and programs work in theory for any number of covariates but in practice we encounter the course of dimensionality (Bellman and Dreyfus, 1962, p. 322) as the number of covariates increase requiring a large sample size to obtain acceptable estimates. A simulation setting is set for the case of two continuous covariates only. In order to check the algorithm performance in larger dimensions we tested the method to estimate functions depending of three and four covariates obtaining good marginal graphs close to the true function.

The purpose of the simulation is to study the performance of the estimators provided by the models described in Section 4.2. We use the reproducing kernel from the thin plate spline regression R_J Section 2.1.1.2 as basis. The arguments in the conclusion from previous chapter, Section 3.4, describe the reasons to use the thin plate splines. In addition, we compare our model that assigns hierarchical normal priors to the latent variables with a model assuming no error in the covariates and takes $\{\hat{x}_i = n^{-1}\sum_{j=1}^{n_w} w_{ij}\}_{i=1}^{n}$ as the true observations of the latent variables. Models with different prior on the variance σ^2, (4.11) – (4.18) are compared as well.

For completeness, we add the model that assumes known σ^2 and uses the true value of this parameter. Table 4.1 summarizes the models to be considered in the simulations.

Under the assumption that the attributes n and σ^2 of the simulated data affect in a similar way the estimation as they did in Chapter 3, the parameters are fixed to $n = 100$ and $\sigma^2 = 0.5^2$ for the rest of this chapter unless stated otherwise. The data was simulated using the form (4.1) and regression function η described by (3.17). The latent variables were generated from $N_2(\mathbf{0}, I_2)$, number of repetitions of the measurement errors range in $n_w \in \{2, 7, 14, 50\}$, the covariance of the measurement errors Σ_w has always in the diagonal the same values varying in $\{0.1^2, 0.1, 0.5^2, 0.5, 1, 2\}$ with correlation 0 in one case, and correlation 0.8 in another case.

For all models we compute the MAPE (3.19) and the SDMAPE (3.20) to measure precision of the regression estimator and the variability around the true function. Furthermore we study the empirical coverage of $C\%$ pointwise credible intervals for predictions and the average empirical coverage (ACP) in the grid $\{\chi_i\}_{i=1}^N$. For the empirical coverage we use $\{\hat{\rho}_i\}_{i=1}^N$ (Section 3.3.3) computed using 150 computations of $C\%$ centered credible intervals from $[\eta(\chi_i)|\mathbf{y}, \mathbf{W}]$; the ACP(C) is estimated as in (3.22).

Figure 4.4 shows an example of the computation of the empirical coverage $\hat{\rho}_i$'s using 150 simulated data sets with $n = 100$, $n_w = 7$ and $\sigma^2 = 0.5^2$; the latent variables $\{\mathbf{x}_i\}_{i=1}^{100}$ are simulated from a bivariate normal standard distribution. The model used to fit the data is a conditional thin plate spline regression, the conditional bandwidth selection method is the RML. The top row is the result of fitting a model that estimates the covariates using a hierarchical normal distribution and inverse-gamma for the observation variance as described in Proposition 11; the bottom row of the figure shows the $\hat{\rho}_i$'s computed with the TPS model from Proposition 6 considering the latent variables observed with values $\left\{n^{-1}\sum_{i=1}^{n_w}\mathbf{w}_{ji}\right\}_{j=1}^{100}$. Each column of the figures is the result of using 35%, 60% or 95% pointwise credible intervals for the predictions. Just from plots in Figure 4.4, observe that the model from the second row has a pointwise empirical coverage and ACP(C) much more lower than the nominal and that the empirical coverage observed in the first row. These effect is more evident looking at the coverage for the rest for the cases.

Table 4.1 Models summary Simulation Study Regression with Measurement Error in the Covariates using Thin Plate Splines. The methods to choose the **Bandwidth** parameter are UERL - unbiased estimate relative loss, GCV - generalized cross validation, RML - restricted maximum likelihood under the Bayes model. Two model specification for the estimation of the latent variables $\hat{\mathbf{x}}$, *Prior* indicates a model with normal Bayes prior as in Proposition 11, *Average* is the model without measurement error case from Chapter 3 using $\{\hat{x}_i = n^{-1}\sum_{j=1}^{n_w} w_{ij}\}_{i=1}^{n}$ as the true covariates. The **priors on σ^2** indicates the models whose priors on the variance are *Normal* as in Proposition 11; *Difference* is the prior (4.12); *Smooth* are the priors (4.17) when bandwidth was chosen with GCV, and prior (4.18) when bandwidth was chosen with RML. For the variance σ^2, *Known* indicates that the model assumes we know and uses σ^2, *Unknown* the model estimates the variance.

Bandwidth	$\hat{\mathbf{x}}$	**Prior σ^2**	σ^2
UERL	Prior	———-	Known
UERL	Prior	Normal	Unknown
UERL	Prior	———-	Known
UERL	Prior	Difference	Unknown
UERL	Prior	———-	Known
UERL	Average	———-	Known
UERL	Average	Normal	Unknown
UERL	Average	———-	Known
UERL	Average	Difference	Unknown
UERL	Average	———-	Known
GCV	Prior	———-	Known
GCV	Prior	Normal	Unknown
GCV	Prior	———-	Known
GCV	Prior	Difference	Unknown
GCV	Prior	———-	Known
GCV	Prior	Smooth	Unknown
GCV	Average	———-	Known
GCV	Average	Normal	Unknown
GCV	Average	———-	Known
GCV	Average	Difference	Unknown
GCV	Average	———-	Known
GCV	Average	Smooth	Unknown
RML	Prior	———-	Known
RML	Prior	Normal	Unknown
RML	Prior	———-	Known
RML	Prior	Difference	Unknown
RML	Prior	———-	Known
RML	Prior	Smooth	Unknown
RML	Average	———-	Known
RML	Average	Normal	Unknown
RML	Average	———-	Known
RML	Average	Difference	Unknown
RML	Average	———-	Known
RML	Average	Smooth	Unknown

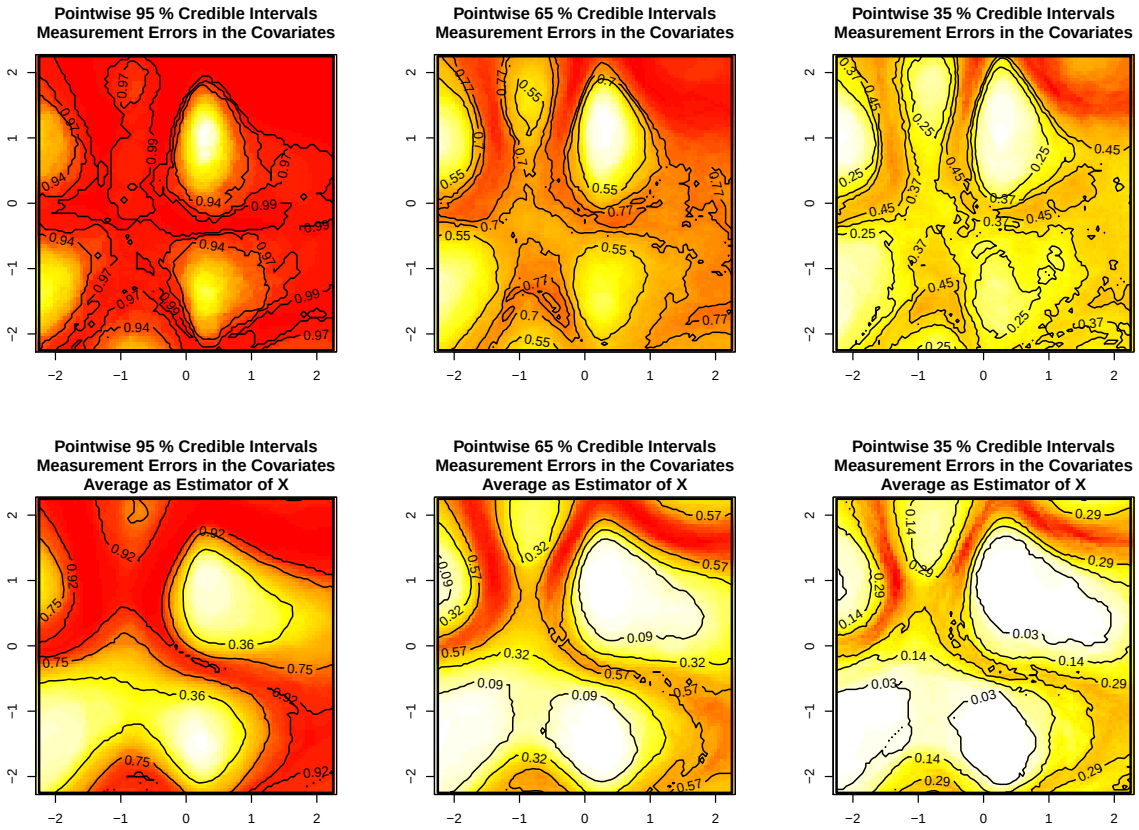

Figure 4.4 Level curves of the empirical coverage for the pointwise 95%, 65% and 35% credible intervals using the Bayesian model with thin plate splines and $m = 3$ for the multivariate regression problem with measurement errors in the covariates. The smoothing parameter is chosen using the restricted maximum likelihood method. The top row of plots are the results of fitting the model that assign a normal prior to the unknown covariates $\{\mathbf{x}_i\}_{i=1}^n$. The bottom row was obtained fitting a model that takes each $\hat{\mathbf{x}}_i = n_w^{-1} \sum_{j=1}^{n_w} \mathbf{w}_{ij}$ and then it is assumed that $\{\mathbf{x}_i\}_{i=1}^n$ are known with $\hat{\mathbf{x}}_i = \mathbf{x}_i$. Each value in the level curves is an estimate of ξ_i for the coverage of $\eta(\chi_i)$ in the grid $\{\chi_i\}_{i=1}^N$ using 100 different simulated data sets and computing the respective credible intervals. Each data set was simulated with $n = 100$, $\sigma^2 = 0.5$ and $\mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$ and $\mathbf{w}_{ij} \overset{iid}{\sim} N_2\left(\mathbf{x}_i, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$. The target function is described by (3.17) and plotted in Figure 3.1.

Table 4.2 summarizes a part of the results from the simulation study and computation of the MAPE and the SDMAPE and their respective 25% and 75% empirical quantiles. Bold numbers for column **25 %** in a given row indicate that the 25% percentile computed with the respective model is larger than the 75% percentile computed with the same assumptions but different way to estimate $\mathbf{x}$; therefore there would be evidence that the median (average by symmetry) MAPE computed over 150 different estimation with the model for that row is larger than the median of the MAPE's computed with the model from row immediately above or below (model that estimate $\mathbf{x}$ different). Bold numbers for column **75 %** indicates that the 75% percentile for the respective combination of parameters is larger than the 25% percentile computed with the same assumptions but different way to estimate $\mathbf{x}$ and the conclusion is similarly as before. Figures 4.5 and 4.6 show boxplots computed using a more comprehensive subset of the simulation results than the ones presented in Table 4.2. These plots show, at the top of the columns, the covariance Σ_w of the measurement errors; each row correspond to estimators using data simulated with n_w repeated measures (proxy variables) of the covariates.

Considering the general estimation and the coverage of the credible intervals, our models provides better estimates and measures of uncertainty as we will describe. But for now, we start by describing the aspect where our model did not provide estimates as good as desired or even worse than a simpler model of taking the averages $\{n_w^{-1}\sum_{j=1}^{n_w}\mathbf{w}_{ij}\}_{j=1}^{n_w}$ as true covariates.

4.4.1 Prediction and variability of prediction for the target regression function and discussion

We start discussing the summary for the MAPE that measures the average absolute difference of the predictions to the true regression function in the grid of prediction $\{\chi_i\}_{i=1}^{N}$. The SDMAPE measures the average variability of the prediction around the true regression.

The most striking feature from Table 4.2, Figure 4.5 and more comprehensive plot in the appendix, Figure C.11, is that the Average MAPE for the model with $\hat{x}_i = \bar{\mathbf{w}}_{i\cdot}$ is smaller than for the model with normal prior on the latent variables when $n_w = 14$ or 50 and even in some cases when $n_w = 7$. This result was indeed not expected for the cases n_w as small as 14 or 7; for $n_w = 50$ Figure C.12 can be easily explained using the law of large numbers, the average

Table 4.2 Summary results partial simulation for $MAPE$ and $SDMAPE$ in the measurement error in the covariates regression. Simulated data with link function (3.17), $n = 100$ and $\sigma^2 = 0.25$. For each latent $\mathbf{x}_i$ are observed $\mathbf{n_w}$ measurement with errors δ_{ij} simulated with $N_2\left(\mathbf{0}, \begin{pmatrix} \sigma_w^2 & 0.8\times\sigma_w^2 \\ 0.8\times\sigma_w^2 & \sigma_w^2 \end{pmatrix}\right)$. **Known σ^2** indicates that the true value of the variance σ^2 was used to estimate the parameters; $\hat{\mathbf{x}}$ indicates the method of estimation of the latent variables. Using 100 repetitions, **Avg M** is the average of the computed $MAPE$, **25 %** and **75 %** are the 25% and 75% empirical percentile of the 100 computed $MAPE$. **Avg SDM** is the average of the 100 computed $SDMAPE$ with their respective 25% and 75% percentiles. Black number for column **25 %** indicate that the 25% percentile computed with the respective model is larger than the 75% percentile computed with the same assumptions but different way to estimate $\mathbf{x}$. Black numbers for column **75 %** indicates that the 75% percentile for the respective combination of parameters is larger than the 25% percentile computed with the same assumptions but different way to estimate $\mathbf{x}$.

$\mathbf{n_w}$	$\sigma_\mathbf{w}$	**Known σ**	$\hat{\mathbf{x}}$	**Avg MAPE**	**25%**	**75%**	**sd MAPE**	**25%**	**75%**
2	1	No	Avg	4.88	4.56	5.25	1.01	0.9	**1.1**
2	1	No	Bayes	4.57	3.91	5	1.50	**1.12**	1.71
2	1	Yes	Avg	4.10	3.74	4.43	0.54	0.47	**0.59**
2	1	Yes	Bayes	3.76	3.11	4.2	1.06	**0.71**	1.19
2	2	No	Avg	5.12	4.68	5.54	1.04	0.96	**1.13**
2	2	No	Bayes	5.92	4.71	7.02	2.24	**1.39**	2.81
2	2	Yes	Avg	4.33	3.96	4.67	0.59	0.53	**0.64**
2	2	Yes	Bayes	5.12	3.86	6.25	1.79	**0.96**	2.43
7	1	No	Avg	2.64	2.27	3.01	0.95	0.86	1.05
7	1	No	Bayes	2.94	2.61	3.18	0.84	0.7	0.93
7	1	Yes	Avg	1.86	1.58	2.16	0.49	0.42	0.56
7	1	Yes	Bayes	2.14	1.89	2.34	0.38	0.27	0.48
7	2	No	Avg	2.78	2.47	**3.12**	0.96	0.87	1.07
7	2	No	Bayes	3.72	**3.23**	4.03	1.12	0.88	1.22
7	2	Yes	Avg	1.99	1.72	**2.22**	0.52	0.44	0.58
7	2	Yes	Bayes	2.91	**2.46**	3.18	0.67	0.44	0.76
14	1	No	Avg	1.97	1.73	**2.24**	0.88	0.8	0.96
14	1	No	Bayes	2.61	**2.4**	2.89	0.73	0.63	0.8
14	1	Yes	Avg	1.16	0.98	**1.39**	0.40	**0.34**	0.47
14	1	Yes	Bayes	1.79	**1.59**	1.93	0.27	0.19	**0.33**
14	2	No	Avg	2.06	1.74	**2.33**	0.92	0.83	1.01
14	2	No	Bayes	2.93	**2.61**	3.19	0.82	0.69	0.92
14	2	Yes	Avg	1.26	0.96	**1.49**	0.47	0.41	0.52
14	2	Yes	Bayes	2.14	**1.86**	2.36	0.38	0.26	0.46
50	1	No	Avg	1.29	1	**1.6**	0.69	0.61	0.75
50	1	No	Bayes	2.13	**1.95**	2.29	0.64	0.55	0.72
50	1	Yes	Avg	0.49	0.28	**0.67**	0.25	0.22	0.29
50	1	Yes	Bayes	1.30	**1.09**	1.46	0.17	0.1	0.23
50	2	No	Avg	1.26	1.06	**1.35**	0.74	0.66	0.8
50	2	No	Bayes	2.31	**2.08**	2.53	0.66	0.6	0.74
50	2	Yes	Avg	0.47	0.22	**0.59**	0.30	0.24	0.33
50	2	Yes	Bayes	1.52	**1.33**	1.63	0.21	0.15	0.25

is becoming a good estimator of the latent variables. With small errors in the covariates, the errors can be ignored and use a model as in Chapter 3.

We cannot fully explain the reasons for the bad results of our proposed model when $n_w = 7$ or 14 as measured by the MAPE. We can only hypothesize using the posterior realizations since $[\mathbf{X}|\mathbf{y}, \mathbf{W}]$ does not have an analytic form. Figure C.13 in the appendix shows a set of estimations of the latent variables $\{\mathbf{x}_i\}_{i=1}^{50}$ for one specific set of simulated data with $n_w = 10$ and $\Sigma_w = \begin{pmatrix} 2 & 0.8\times 2 \\ 0.8\times 2 & 2 \end{pmatrix}$. The yellow-read area represent level curves of the posterior distributions $\{[\mathbf{x}_i|\mathbf{y}, \mathbf{W}]\}_{i=1}^{50}$ using the simulated example. Observe that the estimator provided for each $\mathbf{x}_i$ using the Bayesian model with normal priors on the latent variables has larger variability than the estimator provided using only the averages, $\mathbb{E}(\bar{\mathbf{w}}_{i\cdot}) = \mathbf{x}_i$ and $Var(\bar{\mathbf{w}}_{i\cdot}) = n^{-1}\Sigma_w$. We believe that the differences in the MAPE is because of this variability on the estimation provided by different models.

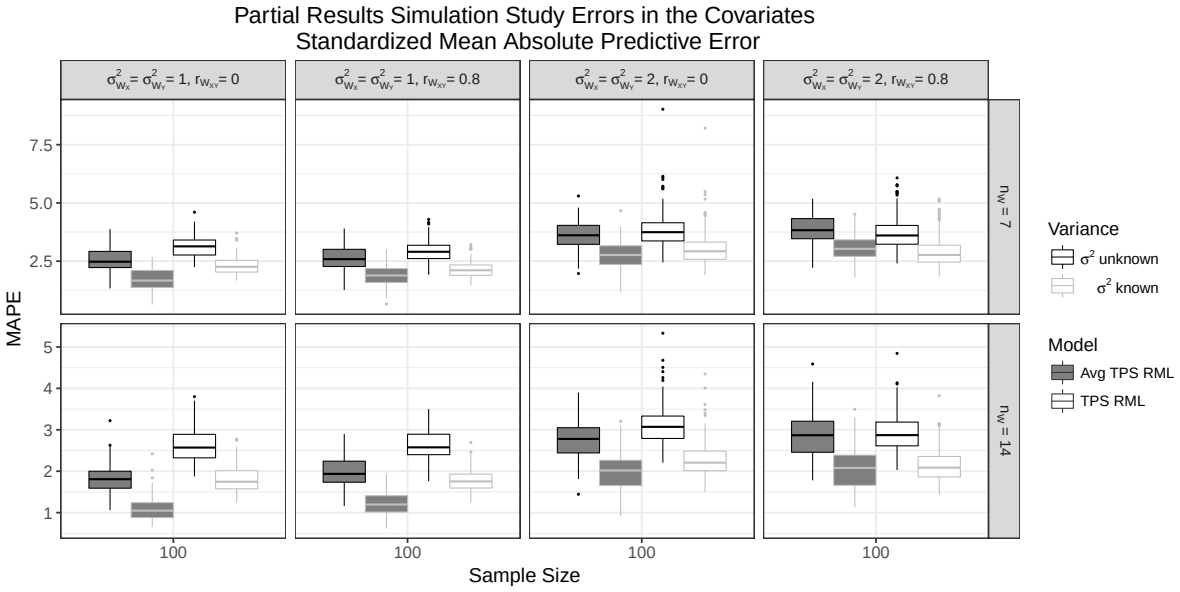

Figure 4.5 Boxplots, part simulation results. Standardized Mean Absolute Predictive Error (3.19) for the multivariate regression problem with measurement errors in the covariates. The MAPE was computed over the square $[-2.25, 2.25]^2$. The content of Table 4.2 is displayed here. The simulated data sets $\{(\mathbf{w}_{i1}, \cdots \mathbf{w}_{in_w}, y_i)\}_{i=1}^{n}$ were generated using the model (4.1), $n = 100$ and $\sigma^2 = 0.5$. The columns in the figure indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the number of repeated observations of the measurement errors of the covariates $\{\mathbf{w}_{j,i}\}_{i=1}^{n_w}$. The models are described in Table 4.1. Complete simulation results in Figure C.11.

By observing a large number of examples as in Figure 4.3 we noticed that the model using the averages as $\hat{\mathbf{x}}$ often produces functions $\hat{\eta}$ that capture the general pattern of the function but misses small features. Furthermore, the function $\tilde{\eta}$, the standard deviation of the posterior predictive distribution, is bounded by a small value which will lead to credible intervals too narrow and to under coverage. For $n_w = 7$ or 14 we observed that the estimates $\hat{\eta}$ with the model on $\mathbf{x}$ as average do not capture any feature similar to the original regression function but in average there is not a significant difference between our model's MAPE and the model with $\mathbf{x}$ as the average.

Nevertheless in practice, with a unique training set provided, the two competing models (using different m if necessary) would produce no significant differences in the predictions in average for $n_w = 7$ or 14 because of the the magnitude of the SDMPAE, see Figures 4.6 and more comprehensively Figure C.14. The SDMAPE summarizes the variability of the estimation around η, the magnitude of the 25% quantiles of the SDMAPE across all simulated cases indicates that in average, the pointwise credible intervals from $[\eta(\chi_i)|\mathbf{y}, \mathbf{W}]$ and different models overlap; in this sense we say that for both available data, both models provide similar estimators. Furthermore, small features from the true η are better described with our model from Proposition 11. For large $n_w = 50$, indeed there is a practical advantage on taking the average estimate as estimates for the latent variables and computing $\hat{\eta}$ in this way, this is true even taking into account the pointwise credible intervals for predicting η.

We have compared prediction and variability for η for models with different priors on the latent variables. Lets' consider now models with differences in the prior for the smoothing parameter λ. There is no statistical difference with respect to prediction and its variability when conditionally choosing the smoothing parameter with the RML method or with the GCV method. We leave open the possibility that the GCV $\mathcal{V}(\lambda, \alpha, \mathbf{X}, \mathbf{y})$ method can be improved through a better calibration of the fudge factor α. We used $\alpha = 1.4$ as suggested trough simulations for the free error cases as explained before. Nevertheless, there is a clear practical and statistical difference when using the $UERL$ score $\mathcal{U}(\lambda, \sigma^2, \mathbf{X}, \mathbf{y})$ with respect to the use of the other two methods; predictions using $\mathcal{U}$ are not to be trusted. The score $\mathcal{U}$ is not independent of σ^2; and as will be described in Section 4.4.2, we are obtaining bad estimates of

the observation error variance σ^2 which may be the cause by inadequate λ and bad predictions for η. We hypothesize that even if we had good estimators of σ^2, the tunning of the smoothing parameter is a sensitive task and the variability of $\hat{\sigma}^2$ can lead to a poor selection of λ and thus, to over-smoothed/under-smoothed $\hat{\eta}$.

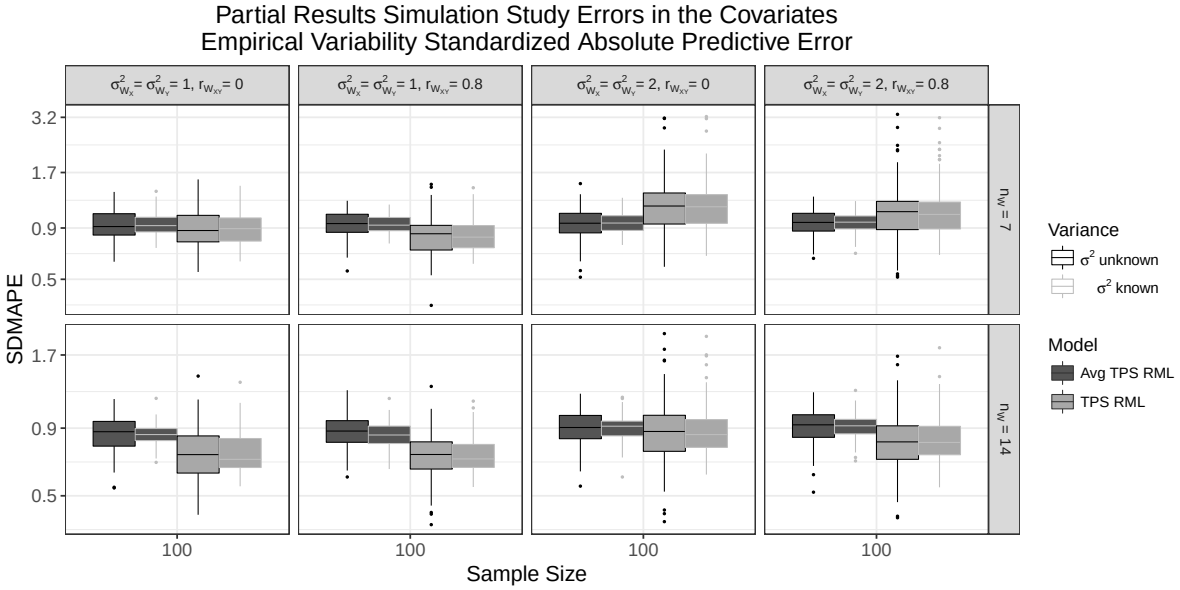

Figure 4.6 Variability Standardized Absolute Predictive Error (3.20) for the multivariate regression problem with measurement errors in the covariates. The SDMAPE was computed over the square $[-2.25, 2.25]^2$. This graphical display contains the information from Table 4.2, last three columns. The simulated data sets $\{(\mathbf{w}_{i1}, \cdots \mathbf{w}_{in_w}, y_i)\}_{i=1}^{100}$ were generated using the model (4.1) and $\sigma^2 = 0.5$. The columns in the figure indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the number of repeated observations of the measurement covariates $\{\mathbf{x}_i\}_{i=1}^{100}$ with errors. The models are described in Table 4.1. Complete simulation results in Figure *C*.14.

We found that there is not a practical and not a significant difference in the average predictions of η for different models that estimate σ^2 with the priors (4.12), (4.17), (4.18) or the inverse gamma as described in the next section. In the next section we compare models with different estimators (priors) for the variance, and found no difference in the estimators for σ^2 and not for the regression function η. But as one can expect, knowing the true value of σ^2 make a practical difference with respect to the prediction of η.

4.4.2 Observation error variance summary results and discussion

The main objective of the models evaluated in this work is to predict the target function. But the full conditional of η requires estimators of the variance of the errors $\{\epsilon\}_{i=1}^{n}$, see model (4.1). The reason we look for more options to estimate the variance is because any effort we made to estimate this parameter was not sufficient. Most of our models over-estimated σ^2, specially when σ^2 is small. We describe our results here.

Proposition 11 describes our main model, such model assigns an inverse gamma prior to σ^2. Modifications on the prior for the variance lead to new models, the modifications we consider are described by equations (4.11), (4.12) and (4.13); such priors use difference frequentist estimators as degenerate priors conditional on the latent variables. Additionally, we considered two more models: the first one uses as prior on λ, the score $\mathcal{V}$ and prior (4.17) on σ^2; the second model uses as prior on λ the score $\mathcal{M}$ and prior (4.18). Table 4.3 summarizes the models we will compare. The comparison we made is not comprehensive but hierarchical, as we will describe.

Table 4.3 Summary of the models based on Proposition 11 varying the priors on the observation error variance σ^2 and the smoothing parameter λ. **Name** indicate the name of the model we will use in the discussions; **Prior on λ (CDF)** indicate the hierarchical prior assigned to the parameters λ, the tail of cumulative distribution function (CDF) is provided; similarly for column **Hierarchical Prior on σ^2 (CDF)**. For the model s_4, $k > 1$.

Name	**Prior on λ (CDF)**	**Hierarchical Prior on σ^2 (upper tail CDF)**
s_0	(4.2), (4.3) or (4.4)	Hierarchical Normal (Proposition 11)
s_1	(4.2), (4.3) or (4.4)	$\int_{\mathbb{R}^n} \mathbf{1}\left\{\sigma^2 \geq \frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i y_{i,1}\right)\right\} dF_{\mathbf{y}\mid\mathbf{X}}(\mathbf{y})$
s_2	(4.2), (4.3) or (4.4)	$\int_{\mathbb{R}^n} \mathbf{1}\left\{\sigma^2 \geq \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_{i,1}\right)^2\right\} dF_{\mathbf{y}\mid\mathbf{X}}(\mathbf{y})$
s_3	(4.2), (4.3) or (4.4)	$\int_{\mathbb{R}^n} \mathbf{1}\left\{\sigma^2 \geq \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - y_{i,1}\right)\left(y_i - y_{i,2}\right)\right\} dF_{\mathbf{y}\mid\mathbf{X}}(\mathbf{y})$
s_4	(4.2), (4.3) or (4.4)	$\int_{\mathbb{R}^n} \mathbf{1}\left\{\sigma^2 \geq \frac{1}{nk^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{k}\left(y_i - y_{i,2,j}\right)\right)\left(\sum_{j=1}^{k}\left(y_i - y_{i,2,j-1}\right)\right)\right\} dF_{\mathbf{y}\mid\mathbf{X}}(\mathbf{y})$
s_5	Distribution (4.3)	$\int_{\mathbb{R}^n} \mathbf{1}\left\{\sigma^2 \geq \sigma_v^2(\lambda, \mathbf{x}, \mathbf{y})\right\} dF_{\mathbf{y}\mid\mathbf{X},\lambda}(\mathbf{y})$
s_6	Distribution (4.4)	$\int_{\mathbb{R}^n} \mathbf{1}\left\{\sigma^2 \geq \sigma_m^2(\lambda, \mathbf{x}, \mathbf{y})\right\} dF_{\mathbf{y}\mid\mathbf{X},\lambda}(\mathbf{y})$

As first step of comparison, we compared models s_1, s_2 and s_3. Such models use the difference frequentist estimators as full conditional posterior for the variance σ^2, see (4.7)−(4.9). We fitted the model described in Proposition 11 with the respective priors and $\mathcal{M}$ in the conditional prior on λ, see (4.4). Given that the objective is to compare estimators of the

variance from these models, the ideal of the cases would be that we knew the regression function; we assumed as well that η is known and is incorporated in the MCMC algorithm used to fit the models. We simulated 200 data sets with the form (4.1), the latent variables were simulated from bivariate standard normal distribution and $n_w \in \{2, 7, 15, 30, 50\}$.

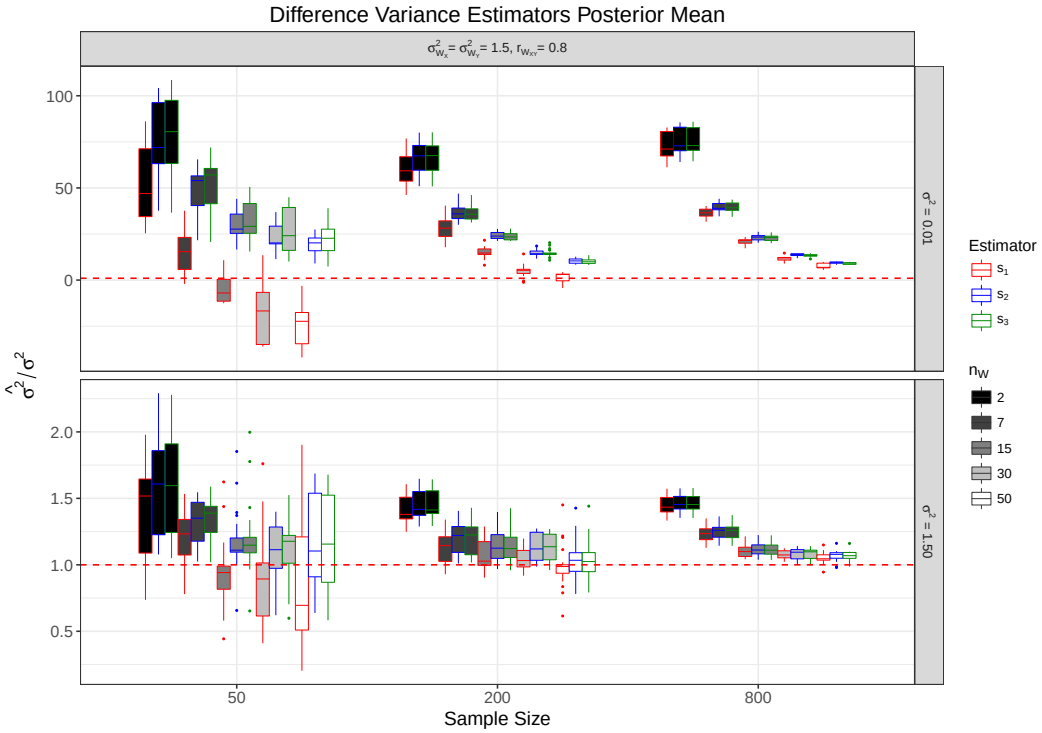

Figure 4.7 Partial results, mean posterior distribution with difference-based estimator as prior on σ^2. Three different degenerated priors on σ^2 identified as *Estimator*. Each estimator is described in Table 4.3. Observe that the vertical axis is different for each plot. The dotted horizontal red line represent the value 1 in the vertical axis.

The mean of the posterior distributions $[\sigma^2|\mathbf{y}, \mathbf{W}, \eta_{true}]$ for each repetition are saved and the result of estimating the variance with these specifications is presented in Figure 4.7 as the ratio between the estimate and the true variance. The desired position of the boxplots is around 1 specified by the red dotted horizontal line. In Figure 4.8 we show the empirical standard deviations of $[\sigma^2|\mathbf{y}, \mathbf{W}]$ for each case and each repetition. Comprehensive results are presented in the appendix plots *C*.17 *C*.18 and *C*.19 and *C*.20.

The most striking feature on these figures is that the mean of the posterior distribution of σ^2 from any model highly overpredicts the true variance, there is large bias. The most

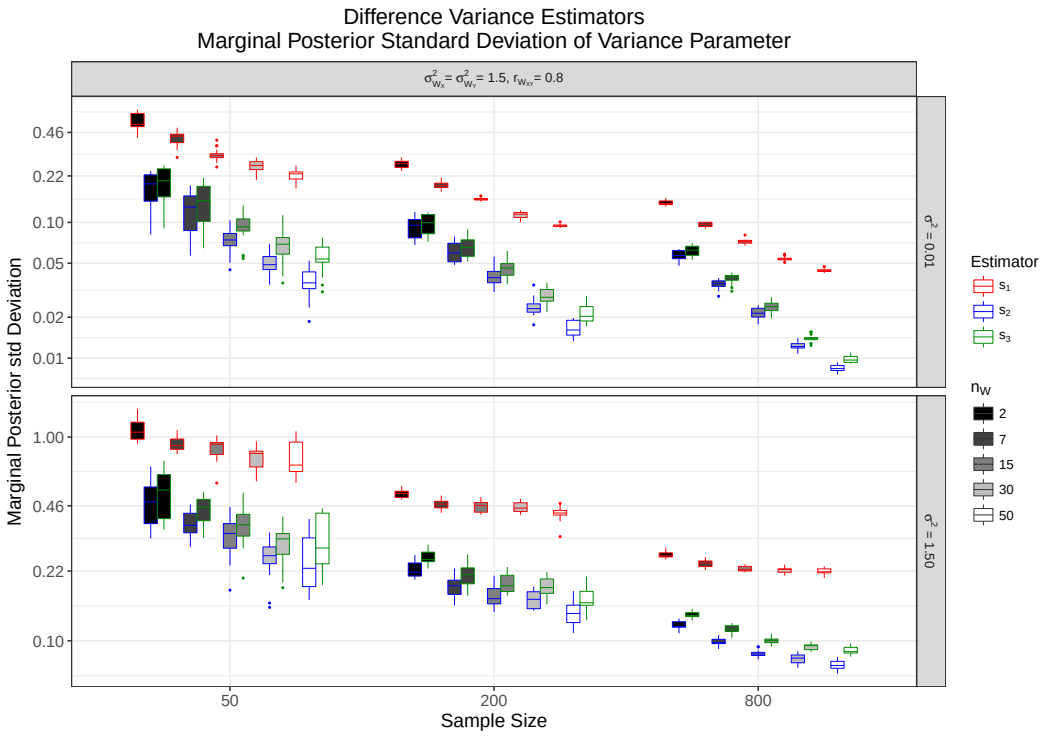

Figure 4.8 Partial results, standard deviation posterior distribution with difference-based estimator as prior on σ^2. Three different degenerated priors on σ^2 identified as *Estimator*. Each estimator is described in Table 4.3. Observe that the vertical axis is different for each plot. The dotted horizontal red line represent the value 1 in the vertical axis.

extreme cases of bias, of a factor of 50 or 75, occur when the variance is small. As the true variance increases, the estimator starts to approximate the true variance but the convergence is slow: when the true variance is 0.5 the factor of error is around 5, and when the true variance $\sigma^2 = 1$ the factor of error has decrease to an average of 1.3. Indeed these methods of estimation need to be improved but it was outside of the direct purpose of this dissertation. In order to choose from these three models, we continue studying the plots. Another feature we observe is that the estimator from model s_1 provides negative values and its variance is the largest for all cases. Comparing models s_2 and s_3 we decide to prefer the former. This decision is based on the magnitude of the bias around σ^2 and the variance of the posterior distribution $[\sigma^2|\mathbf{y}, \mathbf{W}, \eta_{true}]$. We observe that $[\sigma^2|\mathbf{y}, \mathbf{W}, \eta_{true}]$ from model s_2 has a smaller bias and smaller variance in average. We prefer model s_2 above s_3.

Finally, we compare models m_0, m_2, m_5 and m_6. For these cases, we fit the respective models without assuming knownledge of η, but this time we compare the effect of estimating the latent variables with two methods: using hierarchical normal prior as in Proposition 11, assuming $\{n_w^{-1}\sum_{j=1}^{n_w}\mathbf{w}_{i,j}\}_{i=1}^{n}$ as the true latent variables and using model from Chapter 3.

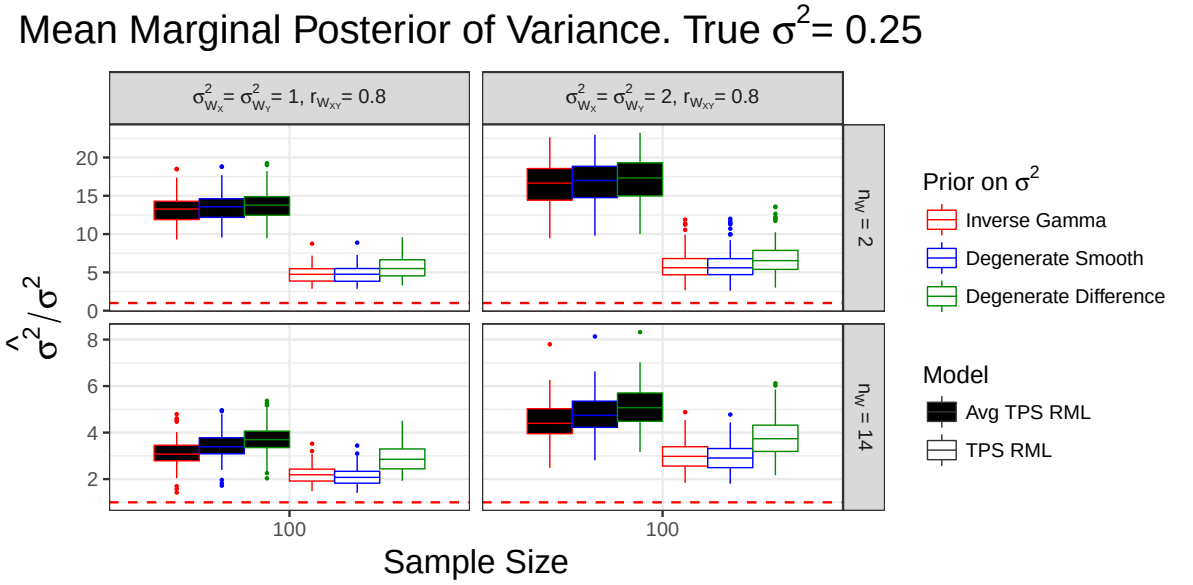

Figure 4.9 Partial simulation results; mean posterior distribution of the observation-error variance σ^2. Observe that the vertical axis is different for different rows. The dotted horizontal red line represent value 1 in the vertical axis. *Inverse Gamma* is the estimator obtained with model m_1, *Degenerated Smooth* is the estimator obtained with model m_6, and *Degenerated Difference* is the estimator obtained with model m_2. Complete simulation results are presented in Figure C.15.

Figure 4.9 summarizes the means of the posterior distributions $[\sigma^2|\mathbf{y},\mathbf{W}]$. Figure C.15 shows a more comprehensive summary of the results. Observe that there is not a significant difference in the means and the posteriors $[\sigma^2|\mathbf{y},\mathbf{W}]$ from these four models when all of them estimate the latent variables with the same method. Nevertheless, as can be expected, estimating the latent variables with the average $\bar{\mathbf{w}}_{i\cdot}$ (*Avg TPS RML* model) does not appropriately take the variability of the errors into account and the estimator of σ^2 is biased by a large factor. The bias in the estimation of σ^2 as observed in this way leads to an under coverage of the $\eta(\chi_i)$ credible intervals. Indeed we have not solved the problem of estimating the variance σ^2 correctly but we have shown numerical evidence that the method of estimation seems to have a rate of

covergence to the true variance that depends on the the number of repeated observations n_w, the covariance of the measurement errors Σ_w and of the true value of σ^2 itself. The convergence rate is slow, even when $n_w = 50$ we have a factor of error of about 2: Figure C.16 shows these cases.

In next section we discuss the empirical coverage of the predictive credible intervals $[\eta(\chi_i)|\mathbf{y}, \mathbf{W}]$. Provided the numerical evidence that there is no difference in estimating the variance σ^2 with either model s_0, s_2, or s_5 and s_6, we decided to use model s_0 to study the coverage of the pointwise credible intervals for prediction of η.

4.4.3 Empirical coverage of credible intervals from predictive posterior distribution for the target regression function and discussion

This section is dedicated to the description of the empirical coverages and the *Across Coverage Probability* ACP(C) for $C\%$ credible intervals obtained through simulation. In a similar fashion as in the free error in the covariates case, for each $\chi_i \in \mathbb{R}^2$ we use the statistics $\{\hat{\rho}_i\}_{i=1}^N$ that measures the empirical coverage of the $C\%$ centered credible intervals from the posterior predictive $[\eta(\chi_i|\mathbf{y}, \mathbf{W})]$ of the respective models. The ACP(C) or $\zeta = \mathbf{E}(\rho_i)$ is estimated using $N^{-1}\sum_{i=1}^N \hat{\rho}_i$. As before, for each combination of parameters, 150 simulated data sets according to (4.1) were generated.

Recall that for the grid $\{\chi_i\}_{i=1}^N$ and a $C\%$ level of credible intervals there is the associate $\{\hat{\rho}_i\}_{i=1}^N$ that can be plotted. Six examples of these plots are shown in Figure 4.4. The distribution of $\{\hat{\rho}_i\}_{i=1}^N$ in each case of combination of parameters, models and C level can be visualized and compared using box plots. Figure 4.10 and its more comprehensive version Figure C.21 show these comparisons. The notation in the plots is the same we have been using so far. The fitted models are described by Proposition 11 and inverse gamma prior on σ^2 was used.

The first noticeable feature of the plots is that the 95% credible intervals provided by the models *Average* are well below the optimal level; for these models, the empirical coverage for most of the evaluations $\{\eta(\chi_i)\}_{i=1}^N$ are below 95%, irrespective of the priors. One would like to have models that produces credible intervals for predictions of $\eta(\chi)$, with χ inside of a convex set containing the latent variables $\{\mathbf{x}_i\}_{i=1}^n$, to have an empirical coverage similar to the optimal

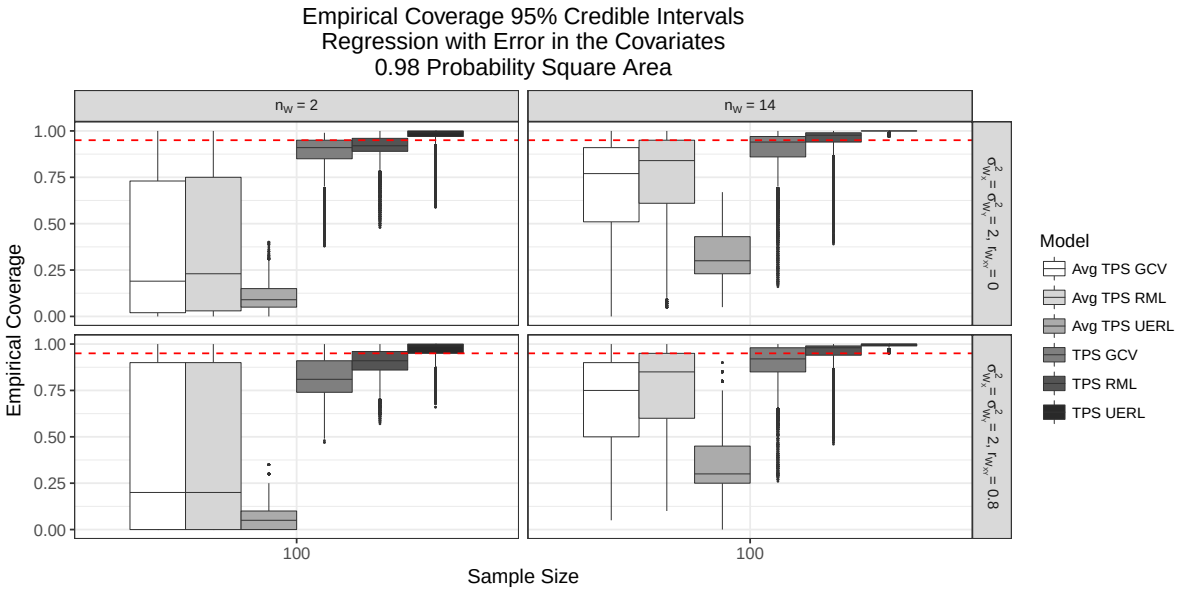

Figure 4.10 Boxplots simulation results, empirical coverage of pointwise 95% credible intervals for prediction of multivariate regression functions with measurement errors in the covariates I. Each box is the summary of the empirical coverages $\{\hat{\rho}_i\}_{i=1}^{N}$. $\hat{\rho}_i \in [0,1]$ is the empirical coverage of the 95% pointwise credible interval for the prediction of $\eta(\chi_i)$ computed after fitting the model to 150 different simulated data sets. The vectors $\{\chi_i\}_{i=1}^{N} \subset \mathbb{R}^2$ is a grid of resolution 0.05×0.05 in the square $[-2.5, 2.5]^2$. The columns in the plot indicate the covariance matrix $\Sigma_W = \begin{pmatrix} \sigma^2_{W_X} & \sigma_{W_X}\sigma_{W_Y} r_{W_{XY}} \\ \sigma_{W_X}\sigma_{W_Y} r_{W_{XY}} & \sigma^2_{W_Y} \end{pmatrix}$ of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated. The models are described in Table 4.1.

level; or at least that the ACP is close to the nominal level, but the *Average* models fail to do so. Only when $n_w = 50$, see Figure C.22, the ACP for the credible intervals of the *Average* models is closer to the nominal but, about 25% of the pointwise credible intervals have an empirical coverage less than 80%. Recall that the UERL method to choose the smoothing parameters depend on estimation of σ^2. Overestimation of this parameter (see Figures 4.9 and C.15) leads to narrower credible intervals, which implies under-coverage of the credible intervals; the under performance is observed even when $n_w = 50$. Similar behavior can be observed in the sequential plots for the average models, Figures 4.11 and more comprehensively in the appendix, see Figures C.23, C.24 and C.25. This relation between overestimation of σ^2 and under coverage of the credible intervals can be explained because for these model, the variance of the process

$[\eta|\mathbf{y}, \mathbf{W}]$ have an inverse proportional relationship with σ^2 trough the full conditional posterior: $Var([\eta(\chi)|\mathbf{y}, \mathbf{W}, \mathbf{X}, \lambda, \sigma^2]) \propto \sigma^{-2}$ (see Proposition 75).

The most striking observations, however, is that the empirical coverage in almost all cases for our model is close to the nominal, at least in the sense of ACP(95). The point-wise coverage of the prediction intervals is fairly close to the nominal with some exceptions. We do not claim that each credible interval has nominal coverage but at least we have the certainty that is within 10% for more than 75% of the intervals. The proximity of the ACP(95) increases as the predictions are computed inside an area of no extrapolation, Figures 4.11, and more comprehensively in C.23, C.24 and C.25. Similar results were found for 60% and 35% credible intervals (not shown here).

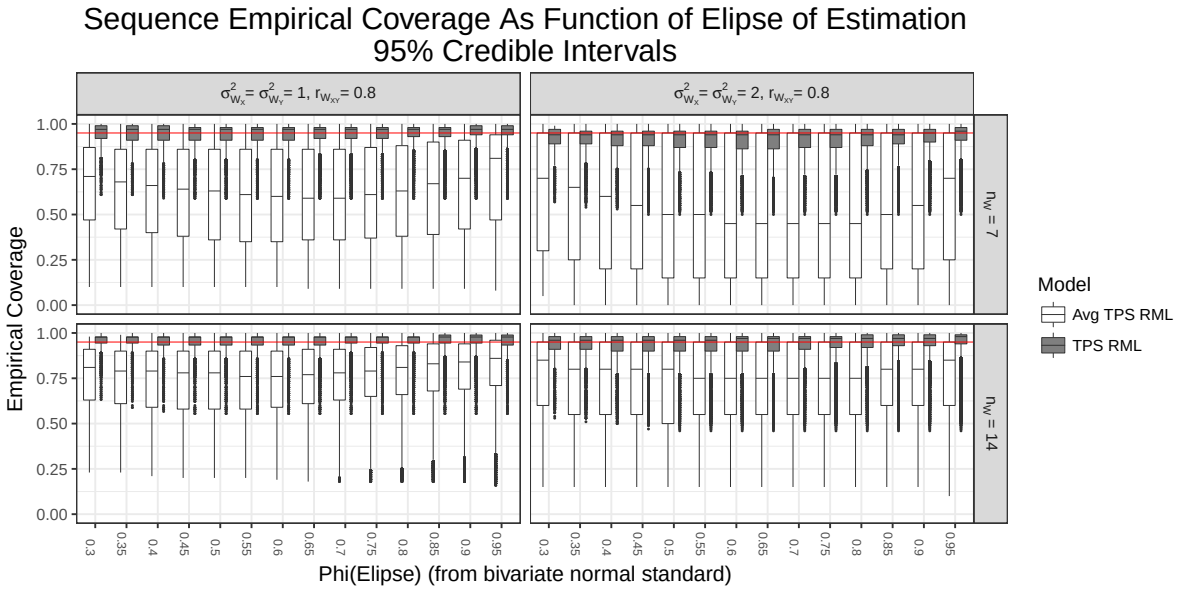

Figure 4.11 Sequential empirical coverage of pointwise 95% credible intervals. The sequence is in the the sense of summarizing the pointwise empirical coverage of the credible intervals $\{\rho_i\}_{i=1}^{m_\alpha}$ for η evaluated in the points $\{\chi_i\}_{i=1}^{m_\alpha}$ from the grid and inside the ellipse that would contain $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution. $alpha = Phi(Ellipse)$. The 150 data sets used to fit iteratively the models and compute the empirical coverage were simulated using $n = 100$, $\sigma^2 = 0.25$, $\mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$ and $\mathbf{w}_{ij} \overset{iid}{\sim} N_2\left(\mathbf{x}_i, \begin{pmatrix} \sigma^2_{W_X} & \sigma_{W_X}\sigma_{W_Y}r_{W_{XY}} \\ \sigma_{W_X}\sigma_{W_Y}r_{W_{XY}} & \sigma^2_{W_Y} \end{pmatrix}\right)$.

4.5 Model Extension: Repeated Responses

We extend the model in Proposition 11 to include multiple observations $\{y_{ij}\}_{j=1}^{n_y}$ for the same latent variable $\mathbf{x}_i$. We assume the available data $\left(\mathbf{w}_{i,1}, \cdots, \mathbf{w}_{i,n_w}, y_{i,1}, \cdots, y_{i,n_y}\right)$, $i = 1, \ldots, n$ with $\mathbf{W}_{i,j} \in \mathbb{R}^d$ and $y_{i,j} \in \mathbb{R}$, can be explained with the model:

$$\begin{aligned} y_{i,j} &= \eta(\mathbf{x}_i) + \epsilon_{i,j} \\ \mathbf{w}_{i,j} &= \mathbf{x}_i + \delta_{i,j} \\ \delta_{i,j} &\sim N_d(\mathbf{0}, \Sigma_w) \\ \epsilon_{i,j} &\overset{iid}{\sim} N(0, \sigma_j^2). \end{aligned} \tag{4.19}$$

If it is not explicitly written, we assume that the parameters are independent. Our aim is to estimate the function η assuming $\eta \in \mathcal{H}^\star$ as in previous Section. The parameters Σ_w and σ_i^2 are estimated as well. A previous work with a similar Bayesian model is described in Castro et al. (2013), but η is assumed to be a linear function by the authors.

Conjecture 12

Let $\mathbb{X}$ a non-empty space, $\mathcal{H}$ a reproducing kernel Hilbert space (RKHS) of functions with domain $\mathbb{X}$ and rank in $\mathbb{R}$. Let J the square norm induced by the semi-inner product of $\mathcal{H}$ which has a null space of finite dimension with basis $\{\psi_i\}_{i=1}^{l}$, let R_J the reproducing kernel of $\mathcal{H}$. Consider a training set of the form $\left(\mathbf{w}_{i,1}, \cdots, \mathbf{w}_{i,n_w}, y_{i,1}, \cdots, y_{i,n_y}\right)$, $i = 1, \cdots, n$ with $\mathbf{w}_{i,j} \in \mathbb{R}^d$ and $y_{i,j} \in \mathbb{R}$. Let $\mathbf{Z} := \{\mathbf{z}_i\}_{i=1}^{k} \subset \{\mathbf{x}_i\}_{i=1}^{n} =: \mathbf{X}$, $\mathbf{d} := (d_1\, d_2\, \cdots\, d_l)^\intercal$, $\mathbf{c} := (c_1\, c_2\, \cdots\, c_k)^\intercal$. Consider the model

$$\begin{aligned} y_{ij} &= \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i) + \epsilon_{ij}, \\ \eta_{\binom{\mathbf{d}}{\mathbf{c}}} &= \sum_{i=1}^{l} d_i \psi_i + \sum_{i=1}^{k} c_i R_J\left(\mathbf{z}_i, \cdot\right) \\ \epsilon_{ij} &\overset{iid}{\sim} N_1(0, \sigma_j^2). \end{aligned}$$

Let $\lambda|\mathbf{X} > 0$, and $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ with entries $Q_{i,j} = R_J(\mathbf{z}_i, \mathbf{z}_j)$, $S|\mathbf{X} \in \mathcal{M}_{n\times l}(\mathbb{R})$ with $S_{i,j}|\mathbf{X} = \psi_j(\mathbf{x}_i)$ full column rank, $R|\mathbf{X} \in \mathcal{M}_{n\times k}(\mathbb{R})$, $R_{i,j}|\mathbf{X} = R_J(\mathbf{x}_i, \mathbf{z}_j)$, $M|\mathbf{X} = RQ^{+}R^\intercal + n\lambda I_n$. Let

$$\tilde{\sigma}^2 = \frac{\prod_{k=1}^{n_y} \sigma_k^2}{\sum_{k=1}^{n_y} \prod_{j\neq k}^{n_y} \sigma_j^2}.$$

Consider the priors

$$d_i \overset{iid}{\sim} 1,$$

$$\mathbf{c}|\tilde{\sigma}^2, \lambda \sim N_k\left(\mathbf{0}, \frac{\tilde{\sigma}^2}{n\lambda} Q^+\right),$$

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X} = \mathbf{x}, \sigma^2 = \tilde{\sigma}^2) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{U}(x|\mathbf{y}, \mathbf{X}, \sigma^2 = \tilde{\sigma}^2)\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y})$$

$$\sigma_i^2 \sim Inv - Gamma(A_\epsilon, B_\epsilon), i = 1, \cdots, n_y$$

$$\mathbf{x}_i|\boldsymbol{\mu}_x, \Sigma_x \overset{iid}{\sim} N_d(\boldsymbol{\mu}_x, \Sigma_x),\ i = 1, .., n,$$

$$\boldsymbol{\mu}_x|\Sigma_x \sim N_d(\mathbf{d}_x, m_x^{-1}\Sigma_x),$$

$$\Sigma_x \sim Inv - Wishart(\mathbf{A}_x, b_x),$$

$$\Sigma_w \sim Inv - Wishart(\mathbf{A}_w, b_w).$$

Unless stated otherwise we consider the parameters to be independent in the priors. Alternatively, we can assign the following conditional priors to λ:

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X} = \mathbf{x}) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{V}(x|\mathbf{y}, \mathbf{X} = \mathbf{x}, \alpha = 1.4)\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}),\ \textit{or}$$

$$\mathbf{P}(\lambda \geq \lambda_0 | \mathbf{X} = \mathbf{x}) = \int_{\mathbb{R}^n} \mathbf{1}\left\{\lambda_0 \geq \arg\min_{x>0} \mathcal{M}(x|\mathbf{y}, \mathbf{X} = \mathbf{x})\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}).$$

Then the joint posterior of the parameters exists and the full conditional posteriors are

- $\binom{\mathbf{d}}{\mathbf{c}}|\mathbf{y}, \sigma_1^2, \cdots, \sigma_{n_y}^2, \lambda, \mathbf{X} \sim N_{l+k}\left(\mu_{\mathbf{dc}}, \frac{\tilde{\sigma}^2}{n\lambda}\boldsymbol{\Sigma}_{\mathbf{dc}}\right)$, *where*

$$\mu_{\mathbf{dc}} = \begin{pmatrix} (S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1} \\ Q^+ R^\intercal M^{-1}(I - S(S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1}) \end{pmatrix} \mathbf{y}$$

$$\boldsymbol{\Sigma}_{\mathbf{dc}} = \begin{pmatrix} (S^\intercal M^{-1} S)^{-1} & -(S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1} R Q^+ \\ -Q^+ R^\intercal M^{-1} S (S^\intercal M^{-1} S)^{-1} & Q^+ - Q^+ R^\intercal \{M^{-1} - M^{-1} S (S^\intercal M^{-1} S)^{-1} S^\intercal M^{-1}\} R Q^+ \end{pmatrix}$$

- $\sigma_i^2|\mathbf{y}, \binom{\mathbf{d}}{\mathbf{c}}, \mathbf{X} \sim Inv - Gamma\left(A_\epsilon + \frac{1}{2}n_y, \left[B_\epsilon^{-1} + \frac{1}{2}\left(\bar{y}_{i,\cdot} - \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i)\right)^2\right]^{-1}\right)$,

-

$$\mathbf{x}_i|\mathbf{y}, \binom{\mathbf{d}}{\mathbf{c}} \propto \prod_{j=1}^{n_y} [y_{i,j}|\sigma_j^2, \mathbf{x}_i] \times \prod_{j=1}^{n_w} [\mathbf{w}_{i,j}|\mathbf{x}_i, \Sigma_w] \times [\mathbf{x}_i|\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{X}}]$$

$$\propto \exp\left[-\frac{1}{2\tilde{\sigma}^2}\left(\frac{\sum_{k=1}^{n_y}\prod_{j\neq k}^{n_y}\sigma_j^2 y_{i,k}}{\sum_{k=1}^{n_y}\prod_{j\neq k}^{n_y}\sigma_j^2} - \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i)\right)^2\right]$$
$$\times \exp\left[-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_x)'\Sigma_x^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_x) - \frac{1}{2}\sum_{j=1}^{n_w}(\mathbf{w}_{ij}-\mathbf{x}_i)'\Sigma_w^{-1}(\mathbf{w}_{ij}-\mathbf{x}_i)\right],$$

- $\Sigma_w|\mathbf{y},\mathbf{x}_1,\cdots,\mathbf{x}_n,\Sigma_x \sim Inv-Wishart\left[\mathbf{A}_w+\boldsymbol{A}\boldsymbol{A}', nn_w+b_w\right]$
  where $\mathbf{A}=[\mathbf{x}_1-\mathbf{w}_{11}..\mathbf{x}_1-\mathbf{w}_{1n_w}\cdots\mathbf{x}_n-\mathbf{w}_{n1}..\mathbf{x}_n-\mathbf{w}_{n,n_w}]$,
- $\boldsymbol{\mu}_x|\mathbf{y},\Sigma_x,\mathbf{x}_1,\cdots,\mathbf{x}_n \sim N_d\left[\Sigma_x^{-1}\left(n\bar{\mathbf{x}}+m_x\mathbf{d}_x\right), \frac{1}{n+m_x}\Sigma_x\right]$, *with* $\bar{\mathbf{x}}=n^{-1}\sum_{i=1}^n\mathbf{x}_i$,
- $\Sigma_x|\mathbf{y},\mathbf{x}_1,\cdots,\mathbf{x}_n = Inv-Wishart\left[\mathbf{A}_x+n\mathbf{S}+\frac{nm_x}{n+m_x}(\bar{\mathbf{x}}-d_x)(\bar{\mathbf{x}}-d_x)', n+b_x\right]$
  where $\mathbf{S}=n^{-1}\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'$,
- $\lambda|\mathbf{y},\tilde{\sigma}^2,\mathbf{X}=\arg\min_{x>0}\left\{\mathcal{U}(x|\tilde{\sigma}^2)\right\}$ *a.s.*,
  If the other priors on λ *depending on* $\mathcal{V}$ *or* $\mathcal{M}$ *were assigned, the full conditional posterior of the parameters do not changes but for* λ. *For* λ *the full conditional posterior distributions are:*

$$\lambda|\mathbf{y},\mathbf{X}=\arg\min_{x>0}\left\{\mathcal{V}(x|\alpha)\right\} \text{ a.s., or}$$
$$\lambda|\mathbf{y},\mathbf{X}=\arg\min_{x>0}\left\{\mathcal{M}(x)\right\} \text{ a.s..}$$

The proof of the existence of the posterior distribution is similar to Proposition 11; it would be required the use of Propositions 72 and 74. The form of the full conditional distributions can be obtained from writing explicitly the posterior and observing the form of the full conditional distribution by completing terms or using the conjugate properties of the gamma inverse distribution, the inverse Wishart distribution.

Furthermore, we could assign degenerated priors to each σ_i^2 as in expressions (4.11) - (4.18) and obtain the respective full conditional posteriors as was described in Section 4.2. A simplification on the assumptions for the conjecture is described in the next corollary.

Corollary 13

In the context of Conjecture 12, if $\sigma_i^2=\sigma^2$ *for* $i=1,...,n_w$. *The posterior exists and the full*

conditional posterior of the parameters are the same as before. Specifically if $\tilde{\sigma}^2 = \sigma^2/n_y$ *then*

$$[\mathbf{x}_i|\Theta\,\mathbf{y}] \propto \exp\left[-\frac{1}{2\tilde{\sigma}^2}\left(\bar{y}_{i,\cdot} - \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i)\right)^2\right]$$
$$\times \exp\left[-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_x)'\Sigma_x^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_x) - \frac{1}{2}\sum_{j=1}^{n_w}(\mathbf{w}_{ij} - \mathbf{x}_i)'\Sigma_w^{-1}(\mathbf{w}_{ij} - \mathbf{x}_i)\right]$$
$$\sigma^2|\mathbf{y},\, \tbinom{\mathbf{d}}{\mathbf{c}},\, \mathbf{X} \sim Inv - Gamma\left(A_\epsilon + \frac{1}{2}n_y n,\, \left[B_\epsilon^{-1} + \sum_{i=1}^{n}\frac{1}{2}\left(\bar{y}_{i,\cdot} - \eta_{\binom{\mathbf{d}}{\mathbf{c}}}(\mathbf{x}_i)\right)^2\right]^{-1}\right).$$

An example of the estimated function η that can be obtained with the models from conjecture 12 using simulated data in the form of (4.19) is presented in Figure 4.12

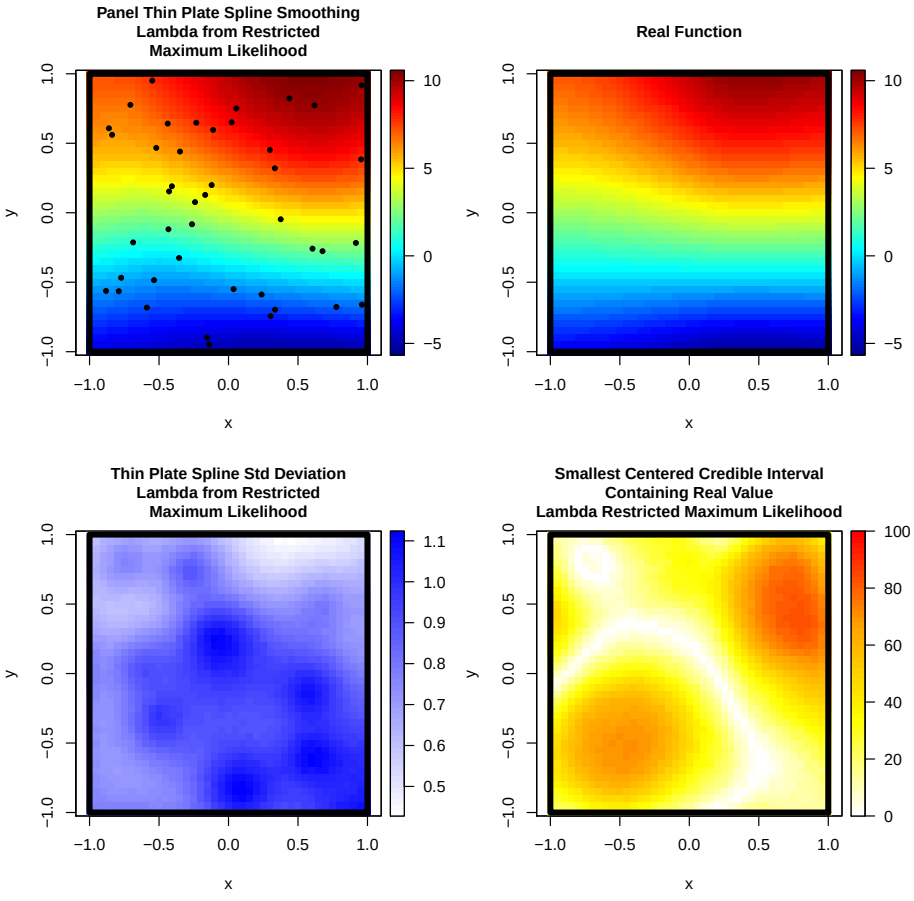

Figure 4.12 Level Curves of η, $\hat{\eta}$, $\tilde{\eta}$ and the minimum level C of a centered pointwise credible interval required to contain the value of $\eta(\chi_i)$. The points in the top left graph are the latent variables. The training data were simulated using model (4.19), η described by (3.17), with $n = 100$, $n_w = 7$, Σ_w diagonal with variances 1. The latent variables were simulated with a standard normal bivariate distribution.

4.6 Conclusions

In the setting of multivariate regression with errors in the covariates and constructing the theory of this chapter using the results from Chapter 2 and 3, we proposed and fully describe Bayesian models to estimate and predict the regression function without directly assuming a form. The Bayesian models use a process to estimate the regression function. The mean of this process has the property that, conditional to the latent variables, solves a regularized least square minimization problem; the mean of the process is obtained as a non-parametric regression from the frequentist context. We use the thin plate splines setting to model the conditional regression function using the machinery of the Bayesian Statistics. In the process of fitting the model we estimate the variance component of the errors as well and any other auxiliary variable used for the estimation process.

Using simulations when the covariates are continuous and in $\mathbb{R}^2$, we found and report the problem of estimating the observation error variance σ^2 and we propose a variety of models to tackle this problem. We introduced different priors on the observation error variance, but still, our first option of inverse-gamma prior was the preferred to estimate σ^2. We conjecture that the elements that affect the rate of convergence of the mean posterior distribution $[\sigma^2\mathbf{y}, \mathbf{W}]$ to σ^2 are n_w, Σ and σ^2. We state that the convergence is slow for $n_w \to \infty$ and specially more difficult to estimate if $\sigma^2 \in \{0.1^2, 0.1, 0.5^2\}$ (and probably for $\sigma^2 \leq 0.5^2$.

We found for $n_w = 50$ (and probably for n_w larger), that estimating the latent variables with the average of its corresponding measures with errors and using this estimation as the true covariates in a usual regression without measurement errors, an *average model*, is enough for practical and efficient predictions of η; nevertheless, the coverage of the credible inervals can be improved for these models. For $n_w \in \{2, 7, 14\}$, (and probably $n_w \in [2, 14]$ or slightly larger interval) the Bayesian model we propose with hierarchical normal distribution for the latent variables provide no practical difference for the point predictors with respect to the η's predictors provided by the *average model*. On the other side, the estimation process of the variance components σ^2 and Σ_w and computation of the credible intervals for predictions require a careful treatment/estimation of the variability of the error in the covariates. The

average model provides biased estimates of the variances and the empirical coverage of the credible intervals for predictions of η is extremely low even in the center region that contain the covariates. Our Bayesian model produce less biased estimates of the variances and the credible intervals have coverage fairly close to the nominal level in the center of the region of observation; the average coverage probability ACP(C) is close to the nominal for most of the regions of estimation. The simulations demonstrate the flexibility of the Bayesian approach, at least in the provided simulations.

Finally, we propose without proof a model to estimate a multivariate regression function in the setting as before but now when repeated observations of the response are available. We conjecture about the full conditional posterior of the the parameters in this model, and we provide an example of the predictions that can be obtained.

We have studied the case of classical error with normal distributions on the errors. We tested the robustness of the model to violations of the normal distribution assumption of the measurement errors $\{\{\delta_{i,j}\}_{j=1}^{n_w}\}_{i=1}^{n}$. We simulated data set with measurement errors with mean zero and distributed as Student, Laplace and Cauchy. We found that the regression function is still recovered.

Certainly, extending the distributions of the measurement errors to other distributions adds additional complexity to the model that would need to be studied. Following these ideas, one can also repeat all the studies and discussion we have developed and assume that the response $\mathbf{y}$ distribution belongs to the exponential family. This new topic is outside the scope of the dissertation but it would be interesting to pursue such research topic.

CHAPTER 5. BAYESIAN MODEL USING THE APPROXIMATED SOLUTION FOR A PENALIZED LEAST SQUARES MINIMIZATION PROBLEM FOR MULTIVARIATE VECTOR VALUED FUNCTIONS

In this Chapter, we focus on learning vector-valued functions using a Bayesan approach that incorporates frequentist results from supervised learning problems where the outputs are vector-valued. The starting idea and assumption of our investigation is that, in practical problems, it is convenient to model the object of interest using functions with multiple outputs and to exploit their dependencies. Thus, even when each output can be modeled separately using the discussions and results from previous chapters, an improvement can be achieved when estimating all the outputs at the same time. Furthermore, we will describe that independent models for each output is a special case of the methods used in this chapter.

Let $\mathbb{X}$ be a non-empty set, let $\mathbb{Y}$ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{Y}}$ and induced norm $\|\cdot\|_{\mathcal{Y}}$, let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subset \mathbb{X} \times \mathbb{Y}$ be an observed sample, let $\mathcal{H} \subset \{\eta : \mathbb{X} \to \mathbb{Y}\}$ be a Hilbert space of functions with semi-inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and induced semi-norm $\|\cdot\|_{\mathcal{H}}$ (Carmeli et al. (2006)). Under these conditions, the regularized functional minimization problem in $\mathcal{H}$

$$\arg\min_{\eta \in \mathcal{H}} \sum_{i=1}^{n} \|\eta(\mathbf{x}_i) - \mathbf{y}_i\|_{\mathbb{Y}}^2 + n\lambda \|\eta\|_{\mathcal{H}}^2 \tag{5.1}$$

is the analogous version of (2.1) for vector valued functions. The solution to this minimization problem has only been recently studied using the perspective of linear operators in vector valued Hilbert spaces $\mathcal{H}$ (Micchelli and Pontil (2004); Carmeli et al. (2006); Caponnetto et al. (2008); Agarwal et al. (2010)). As in the univariate case, the Representer Theorem for vector valued functions (Carmeli et al. (2006); Caponnetto and De Vito (2007); Minh et al. (2013)) provides sufficient conditions for the existence and for the form of the solution to (5.1), but in most of the cases, it is not feasible to compute the required expressions for an explicit solution to (5.1).

Nevertheless, in the case that $\mathcal{H} \subset \{\eta : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}\}$, the solution can be, in theory, explicitly computed by solving a large system of linear equations. In this chapter we propose a Bayesian approach to solve the non parametric regression problem using an approximate solution to the functional minimization problem (5.1). Such approximation appears as the mean of a random process defined by one of the full conditional posterior distributions.

The minimal needed theory of *RKHS* of vector valued functions is revisited in Section 5.1. In Section 5.2 we propose and describe a Bayesian approach to estimate multivariate vector-valued regression functions η in a Hilbert space without assumptions on the form. Furthermore, we interpret the proposed models in the context of a frequentist non parametric regression problems. We propose three algorithms to select a real valued bandwidth parameters. Some of their properties are conjectured and we further theoretically address calculations of bandwidth diagonal matrices and complete bandwidth matrices. We provide details on the fitting method of the models. Finally, in Section 5.3 we extend the the current Bayesian approach and propose Bayesian estimators for vector-valued functions when the data available in the regression problem is contaminated with measurement errors in the covariates in classical sense. We illustrate the methods with some examples.

5.1 Preliminaries

Let $\mathbb{X}$ be a non-empty set, let $\mathbb{Y}$ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{Y}}$ and induced norm $\|\cdot\|_{\mathbb{Y}}$. Let $\mathfrak{L}(\mathbb{Y})$ be the Banach space (Definition 36) of bounded linear operators on $\mathbb{Y}$.

Definition 14

Functions $\mathcal{K} : \mathbb{X} \times \mathbb{X} \to \mathfrak{L}(\mathbb{Y})$ *are positive definite kernels if for any* $(\mathbf{x}, \mathbf{z}) \in \mathbb{X} \times \mathbb{X}$, $\mathcal{K}(\mathbf{x}, \mathbf{z}) \in \mathfrak{L}(\mathbb{Y})$ *is a self adjoint operator (Definition 37) and*

$$0 \leq \sum_{i,j=1}^{n} \langle \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)\mathbf{y}_i, \mathbf{y}_j \rangle_{\mathbb{Y}}\,, \tag{5.2}$$

for every finite sets $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{X}$ *and* $\{\mathbf{y}_i\}_{i=1}^n \subset \mathbb{Y}$.

Remark 15

If $\mathbb{Y} = \mathbb{R}^d$, the family of self adjoint operators are represented by the space of semipositive definite matrices $\mathcal{M}^{+}_{d_2 \times d_2}(\mathbb{R})$.

Notation 16 *Let $\mathcal{M}^{++}_{d_2 \times d_2}(\mathbb{R})$ denote the positive definite matrices.*

Given a positive definite kernel $\mathcal{K}$ in the context we have discussed, there exist a unique $\mathbb{Y} - valued\ RKHS\ \mathcal{H}_{\mathcal{K}}$ that has $\mathcal{K}$ as its reproducing kernel (Carmeli et al., 2006, Proposition 2.3). Let $\langle \cdot, \cdot \rangle_{\mathcal{H}_{\mathcal{K}}}$ be the inner product and $\|\cdot\|_{\mathcal{H}_{\mathcal{K}}}$ the induced norm. The *reproducing property* satisfied by $\mathcal{K}$ in $\mathcal{H}_{\mathcal{K}}$ is

$$\langle \eta, \mathcal{K}(\cdot, \mathbf{x})\mathbf{y} \rangle_{\mathcal{H}_{\mathcal{K}}} = \langle \eta(\mathbf{x}), \mathbf{y} \rangle_{\mathbb{Y}}, \text{ for all } \eta \in \mathcal{H}_{\mathcal{K}}. \tag{5.3}$$

.

Furthermore, let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subset \mathbb{X} \times \mathbb{Y}$ a labeled training set, and let $\mathbb{Y}$ be a separable Hilbert space (Definition 38) then, the Representer Theorem 81 and 82 state that the solution to (5.1) is unique and have the form

$$\eta(\mathbf{x}) = \sum_{i=1}^{n} \mathcal{K}(\mathbf{x}_i, \mathbf{x})\mathbf{a}_i. \tag{5.4}$$

In a similar way as in Section 2.2 we propose to approximate the solution to (5.1) by minimizing in a subspace $\mathcal{H}^{\star}_{\mathcal{K}} \subsetneq \mathcal{H}_{\mathcal{K}}$. For a given set of knots $\{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n$, define $\mathcal{H}^{\star}_{\mathcal{K}} = span\{\mathcal{K}(\mathbf{z}_i, \cdot), i = 1, ..., k\}$. The number of knots k and the knots $\{\mathbf{z}_i\}_{i=1}^k$ can be chosen as in Section 2.2. Therefore, the solution to (5.1) in the space $\mathcal{H}^{\star}_{\mathcal{K}}$ has the form

$$\eta^{\star}(\mathbf{x}) = \sum_{i=1}^{k} \mathcal{K}(\mathbf{z}_i, \mathbf{x})\mathbf{a}_i. \tag{5.5}$$

The space $\mathcal{H}^{\star}_{\mathcal{K}}$ is clearly a reproducing kernel Hilbert space with reproducing kernel $\mathcal{K}$. Thus, by Theorem 81 and 82 there exist a unique solution $\hat{\eta^{\star}}$ to the minimization problem in this space. Let us denote with $\hat{\eta}$ to the solution of (5.1) in $\mathcal{H}$, and $\hat{\eta}^{\star}$ to the solution of (5.1) in $\mathcal{H}^{\star}$, then it is not difficult to see that $\hat{\eta}^{\star} \overset{k \to n}{\longrightarrow} \hat{\eta}$ because $\hat{\eta} \in \mathcal{H}^{\star}$ as $k \to n$, $\mathcal{H}^{\star} \subset \mathcal{H}$ for any k and because the solution to (5.1) is unique.

We do not claim that the convergence properties of the approximated solution $\hat{\eta}^{\star}$ discussed for the real response regression problem in Section 2.2 hold for the current setting, but we

propose $\hat{\eta}^\star$, $k << n$, as approximation to the exact solution of (5.1) because of computational feasibility at the moment of fitting the models. Furthermore, the proposition of using $\hat{\eta}^\star$ as estimator of the true function η in the hope that we do not loose too much with respect to the full solution $\hat{\eta}$.

Now, lets consider a special case of Hilbert space and explicitly compute the solution to the minimization problem in $\mathcal{H}_{\mathcal{K}}^\star$. Let $\mathbb{X} = \mathbb{R}^{d_1}$, and let $\mathbb{Y} = \mathbb{R}^{d_2}$ the Hilbert space with inner product $\langle \mathbf{x}, \mathbf{z} \rangle_{\mathbb{Y}} = \mathbf{x}^\intercal \Sigma^{-1} \mathbf{z}$ and $\Sigma \in \mathcal{M}_{d_2 \times d_2}^{++}(\mathbb{R})$.

Remark 17

The fact that $\mathbb{Y}$ with the above structure is a Hilbert space would need to be proven, but this follow immediately from the definition of Hilbert space and that the quadratic form $\mathbf{x}^\intercal \Sigma^{-1} \mathbf{z}$ is an inner product. The quadratic form is an inner product because $\Sigma \in \mathcal{M}_{d_2 \times d_2}^{++}(\mathbb{R})$ (Banerjee and Roy (2014)).

Plugging (5.5) in the functional

$$\mathfrak{L}(\eta) = \sum_{i=1}^{n} (\mathbf{y}_i - \eta(\mathbf{x}_i))^\intercal \Sigma^{-1} (\mathbf{y}_i - \eta(\mathbf{x}_i)) + n\lambda \|\eta\|_{\mathcal{H}_{\mathcal{K}}}^2$$

and using the reproducing properties of $\mathcal{K}$ (Proposition 77) we obtain that $\mathfrak{L}$ restricted to the space $\mathcal{H}_{\mathcal{K}}^\star$ can be written as follow

$$\mathfrak{J}(\mathcal{A}) := \mathfrak{L}_{|\mathcal{H}_{\mathcal{K}}^\star}(\eta) = (\mathbf{Y} - \mathbf{K}_{xz}\mathcal{A})^\intercal \Psi_{\Sigma,n} (\mathbf{Y} - \mathbf{K}_{xz}\mathcal{A}) + n\lambda \mathcal{A}^\intercal \mathbf{\Gamma} \mathcal{A}, \tag{5.6}$$

where $\mathbf{Y} = \text{vec}(\mathbf{y}_1 \cdots \mathbf{y}_n)$, $\mathcal{A} = \text{vec}(\mathbf{a}_1 \cdots \mathbf{a}_k)$, $\mathbf{K}_{xz} \in \mathcal{M}_{nd_2 \times kd_2}(\mathbb{R})$ is a block matrix with i,jth block $\mathcal{K}(\mathbf{x}_i, \mathbf{z}_j)$, $\mathbf{K}_{zz} \in \mathcal{M}_{kd_2 \times xd_2}(\mathbb{R})$ is another block matrix with i,jth block $\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)$, $\mathbf{K}_{zx} = \mathbf{K}_{xz}^\intercal$, $\Psi_{\Sigma,n} = I_n \otimes \Sigma^{-1}$ and $\mathbf{\Gamma} = \mathbf{K}_{zz}\Psi_{\Sigma,n} + \Psi_{\Sigma,n}\mathbf{K}_{zz}$. The problem of solving (5.1) in the space $\mathcal{H}_{\mathcal{K}}^\star$ is reduced to the minimization of (5.6) in $\mathbb{R}^{kd_2}$.

By Proposition 77 the critical points of $\mathfrak{L}$ satisfy the system of equations

$$\{\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + n\lambda\mathbf{\Gamma}\}\mathcal{A} = \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{Y}, \tag{5.7}$$

and furthermore, Proposition 77 states that any solution to (5.7) leads to the computation of the global maximum of $\mathfrak{L}$ by using (5.5) jointly with any critical point of $\mathfrak{J}$. This property will be helpful to further prove properties of our proposed model in Section 5.2. There, we propose

a Bayesian model such that mean of the full conditional posterior of an induced process, solves the minimization problem 5.1 in $\mathcal{H}^\star$. We address the selection of an smoothing parameter $\lambda > 0$ and theoretically extend to selection of bandwidth matrices in Section 5.2.2. For now, we assume we are given the smoothing parameter related to the minimization problem (5.1) $\lambda > 0$.

5.2 Bayes Regression Models

We have everything ready to state one of our main propositions for the regression problem with observed training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, $\{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n$ and $\lambda > 0$. Consider the model

$$\begin{aligned} \mathbf{y}_i &= \eta(\mathbf{x}_i) + \epsilon_i \\ \epsilon_i &\overset{iid}{\sim} N_{d_2}(\mathbf{0}, \Sigma). \end{aligned} \tag{5.8}$$

Of particular interest is to estimate η using a nonparametric regression method, but from Section 5.1, assuming a form as (5.5) with a chosen reproducing kernel $\mathcal{K}$ is justified. For the form assumed of η, the unknown parameters $\mathcal{A} := \operatorname{vec}(\mathbf{a}_1 \cdots \mathbf{a}_k)$, $\{\mathbf{a}_i\}_{i=1}^k \subset \mathbb{R}^{d_2}$ and $\Sigma \in \mathcal{M}_{d_2 \times d_2}^{++}(\mathbb{R})$ are to be estimated. All such ideas can be incorporated in a Bayesian model. According to the Bayesian theory, we can provide Bayes estimators of Σ, predictors of η as a process, and credible intervals for $\eta(\chi)$, $\chi \in \mathbb{R}^{d_1}$.

5.2.1 A first Bayes regression model

The next Proposition describes one of the Bayesian models we propose. A smoothing parameter $\lambda > 0$ and a reproducing kernel $\mathcal{K}$ are assumed to be given.

Proposition 18

Let $\mathcal{H}$ be a RKHS of functions with domain $\mathbb{R}^{d_1}$ and rank in $\mathbb{R}^{d_2}$, let $\mathcal{K}$ be the reproducing kernel of $\mathcal{H}$, let the pairs $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ be an observed labeled training set, let $\mathbf{Z} := \{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n =: \mathbf{X}$ be the knots, let $\lambda > 0$ be the smoothing parameter, let $b = \frac{2}{n\lambda}$, $\mathbf{Y} = \operatorname{vec}(\mathbf{y}_1 \, \mathbf{y}_2 \cdots \mathbf{y}_n)$ observed vector-valued response, let $\mathcal{A} = \operatorname{vec}(\mathbf{a}_1 \, \mathbf{a}_2 \cdots \mathbf{a}_k)$, $\mathbf{a}_i \in \mathbb{R}^{d_2}$ be the coefficients to be estimated. The auxiliary matrices are defined as $\mathbf{K}_{xz} \in \mathcal{M}_{nd_2 \times kd_2}(\mathbb{R})$ a

block matrix with (i,j)th block $\mathcal{K}(\mathbf{x}_i,\mathbf{z}_j)$, $\mathbf{K}_{zz}\in\mathcal{M}_{kd_2\times xd_2}(\mathbb{R})$ another block matrix with (i,j)th block $\mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)$, $\mathbf{K}_{zx}=\mathbf{K}_{xz}^{\intercal}$, $\Psi_{\Sigma,n}=\left(I_n\otimes\Sigma^{-1}\right)$, $\mathbf{\Gamma}=\Psi_{\Sigma,k}\mathbf{K}_{zz}+\mathbf{K}_{zz}\Psi_{\Sigma,k}$. Consider the model

$$\begin{aligned}\mathbf{y}_i&=\eta(\mathbf{x}_i)+\epsilon_i,\\ \eta(\mathbf{x})&=\sum_{i=1}^{k}\mathcal{K}(\mathbf{z}_i,\mathbf{x})\mathbf{a}_i,\\ \epsilon_i&\overset{iid}{\sim}N_{d_2}(\mathbf{0},\Sigma).\end{aligned}$$

Consider the priors on the parameters

$$\begin{aligned}\mathcal{A}|\Sigma&\sim N_{kd_2}\left[\mathbf{0},b\mathbf{\Gamma}^{+}\right]\\ \Sigma&\sim Inv-Wishart(\mathbf{A},\nu),\end{aligned}$$

where $\mathbf{\Gamma}^{+}$ is the Moore-Penrose inverse of the matrix $\mathbf{\Gamma}$. Then the posterior of the parameters exists and the full conditional posteriors are

$$\begin{aligned}\Sigma|\mathbf{Y},\mathcal{A}\sim&Inv-Wishart\left(\Sigma_p,\nu+k+1\right),\\ \mathcal{A}|\mathbf{Y},\Sigma\sim&N_{kd_2}\left[\mu_{\mathbf{Y}},\Sigma_{\mathbf{Y}}\right],\\ \mu_{\mathbf{Y}}=&\mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx}+\frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}\mathbf{Y}\\ \Sigma_{\mathbf{Y}}=&b\left\{\mathbf{\Gamma}^{+}-\mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx}+\frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\right\}\\ \Sigma_p=&\mathbf{A}+n\lambda\sum_{i=1}^{k}[\mathcal{A}\mathcal{A}^{\intercal}\mathbf{K}_{zz}]_{(i)(i)}\\ &+\begin{pmatrix}\mathbf{y}_1-\eta(\mathbf{x}_1)&\cdots&\mathbf{y}_n-\eta(\mathbf{x}_n)\end{pmatrix}\begin{pmatrix}(\mathbf{y}_1-\eta(\mathbf{x}_1))^{\intercal}\\ \vdots\\ (\mathbf{y}_n-\eta(\mathbf{x}_n))^{\intercal}\end{pmatrix},\end{aligned}$$

where $[\mathcal{A}\mathcal{A}^{\intercal}\mathbf{K}_{zz}]_{(i)(i)}\in\mathcal{M}_{d_2\times d_2}(\mathbb{R})$ is the (i,i)th block matrix in the diagonal of $\mathcal{A}\mathcal{A}^{\intercal}\mathbf{K}_{zz}$.

Proof.

We need to proof first that the priors are well defined. In particular we need to prove that $\mathbf{\Gamma}^{+}$ is a valid covariance matrix. It is trivial that $\mathbf{\Gamma}$ is symmetric. The semipositive definite property required for $\mathbf{\Gamma}^{+}$ can be proven using the square norm $\|\eta\|^2_{\mathcal{H}_{\mathcal{K}}}$ and the properties

$\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j) = \mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)^\intercal = \mathcal{K}(\mathbf{z}_j, \mathbf{z}_i)^\intercal = \mathcal{K}(\mathbf{z}_j, \mathbf{z}_i)$ and that $\mathcal{K}$ is a reproducing kernel, as we do now. Let $\{\mathbf{a}_i\}_{i=1}^k \subset \mathbb{R}^{d_2}$ any subset of vectors, let $\eta = \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i, \cdot)\mathbf{a}_i$, then

$$
\begin{aligned}
0 \leq \|\eta\|^2_{\mathcal{H}_\mathcal{K}} &= \left\langle \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i, \cdot)\mathbf{a}_i, \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i, \cdot)\mathbf{a}_i \right\rangle_{\mathcal{H}_\mathcal{K}} \\
&= \sum_{i,j=1}^k \langle \mathcal{K}(\mathbf{z}_i, \cdot)\mathbf{a}_i, \mathcal{K}(\mathbf{z}_j, \cdot)\mathbf{a}_j \rangle_{\mathcal{H}_\mathcal{K}} \\
&= \frac{1}{2}\left(\sum_{i,j=1}^k \langle \mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\mathbf{a}_i, \mathbf{a}_j \rangle_{\mathbb{R}^{d_2}} + \sum_{i,j=1}^k \langle \mathbf{a}_i, \mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\mathbf{a}_j \rangle_{\mathbb{R}^{d_2}} \right) \qquad (5.9) \\
&= \frac{1}{2}\left(\sum_{i,j=1}^k \mathbf{a}_i^\intercal \mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\Sigma^{-1}\mathbf{a}_j + \sum_{i,j=1}^k \mathbf{a}_i^\intercal \Sigma^{-1}\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\mathbf{a}_j \right) \\
&= \frac{1}{2}\sum_{i,j=1}^k \mathbf{a}_i^\intercal \left\{ \mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\Sigma^{-1} + \Sigma^{-1}\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j) \right\} \mathbf{a}_j \\
&= \frac{1}{2}\mathcal{A}^\intercal \left(\mathbf{K}_{zz}\left(I_k \otimes \Sigma^{-1}\right) + \left(I_k \otimes \Sigma^{-1}\right)\mathbf{K}_{zz} \right) \mathcal{A}, \\
&= \mathbf{\Gamma}. \qquad (5.10)
\end{aligned}
$$

where $\mathcal{A}$ as defined in the statement of the Theorem and (5.9) was obtained using the reproducing property of $\mathcal{K}$ and symmetry of the inner product $\langle \cdot, \cdot \rangle_{\mathbb{R}^{d_2}}$. Thus $\mathbf{\Gamma}$ is semi positive definite and $\mathbf{\Gamma}^+$ is semi positive definite as well.

The joint posterior of the parameters exists because all the priors are proper. Such posterior distribution can be written as follow

$$
[\mathcal{A},\, \lambda,\, \Sigma | \mathbf{Y},\, \mathbf{X}] = [\mathbf{Y} | \mathcal{A},\, \mathbf{X},\, \lambda,\, \Sigma] \times [\mathcal{A} | \Sigma,\, \mathbf{X},\, \lambda] \times [\Sigma | \mathbf{X},\, \lambda] = [\mathbf{Y},\, \mathcal{A} | \mathbf{X},\, \lambda,\, \Sigma] \times [\Sigma]\,.
$$

First we compute the full conditional posterior $[\mathcal{A} | \Sigma,\, \mathbf{X},\, \lambda]$ by observing that $[\mathbf{Y},\, \mathcal{A} | \mathbf{X},\, \lambda,\, \Sigma]$ is a multivariate normal distribution. Let's compute its mean and covariance matrix.

$$
\begin{aligned}
\mathbb{E}\left(\mathbf{y}_i \middle|\, \mathbf{X}, \lambda, \Sigma\right) &= \mathbb{E}\left(\sum_{j=1}^k \mathcal{K}(\mathbf{z}_j, \mathbf{x}_i)\mathbf{a}_j + \epsilon_i \middle|\, \mathbf{X}, \lambda, \Sigma \right) \\
&= \sum_{j=1}^k \mathcal{K}(\mathbf{z}_j, \mathbf{x}_i)\mathbb{E}\left(\mathbf{a}_j \middle|\, \mathbf{X}, \lambda, \Sigma\right) + \mathbb{E}\left(\epsilon_i \middle|\, \mathbf{X}, \lambda, \Sigma\right) \\
&= \mathbf{0},
\end{aligned}
$$

$$\mathbb{E}\left(\mathbf{a}_i \middle| \mathbf{X}, \lambda, \Sigma\right) = \mathbf{0},$$

and

$$\begin{aligned}
Cov\left[\mathbf{y}_i, \mathbf{y}_j \middle| \mathbf{X}, \lambda, \Sigma\right] &= Cov\left[\sum_{o=1}^{k} \mathcal{K}(\mathbf{z}_o, \mathbf{x}_i)\mathbf{a}_o + \epsilon_i, \sum_{\nu=1}^{k} \mathcal{K}(\mathbf{z}_\nu, \mathbf{x}_j)\mathbf{a}_\nu + \epsilon_j \middle| \mathbf{X}, \lambda, \Sigma\right] \\
&= \sum_{o=1}^{k}\sum_{\nu=1}^{k} Cov\left[\mathcal{K}(\mathbf{z}_o, \mathbf{x}_i)\mathbf{a}_o, \mathcal{K}(\mathbf{z}_\nu, \mathbf{x}_j)\mathbf{a}_\nu \middle| \mathbf{X}, \lambda, \Sigma\right] + \mathbf{1}_{\{i=j\}}\Sigma \\
&= \sum_{o=1}^{k}\sum_{\nu=1}^{k} \mathcal{K}(\mathbf{z}_o, \mathbf{x}_i) Cov(\mathbf{a}_o, \mathbf{a}_\nu)\mathcal{K}(\mathbf{z}_\nu, \mathbf{x}_i) + \mathbf{1}_{\{i=j\}}\Sigma \\
&= b\left[\mathcal{K}(\mathbf{x}_i, \mathbf{z}_1) \cdots \mathcal{K}(\mathbf{x}_i, \mathbf{z}_k)\right] \mathbf{\Gamma}^{+} \begin{bmatrix} \mathcal{K}(\mathbf{x}_j, \mathbf{z}_1) \\ \vdots \\ \mathcal{K}(\mathbf{x}_j, \mathbf{z}_k) \end{bmatrix} + \mathbf{1}_{\{i=j\}}\Sigma, \\
Cov\left[\mathbf{a}_i, \mathbf{y}_j \middle| \mathbf{X}, \lambda, \Sigma\right] &= Cov\left[\mathbf{a}_i, \sum_{\nu=1}^{k} \mathcal{K}(\mathbf{z}_\nu, \mathbf{x}_j)\mathbf{a}_\nu + \epsilon_j \middle| \mathbf{X}, \lambda, \Sigma\right] \\
&= \sum_{\nu=1}^{k} Cov(\mathbf{a}_i, \mathbf{a}_\nu)\mathcal{K}(\mathbf{z}_\nu, \mathbf{x}_j) \\
&= b\left[\mathbf{\Gamma}^{+}_{(i)(1)}\, \mathbf{\Gamma}^{+}_{(i)(2)} \cdots \mathbf{\Gamma}^{+}_{(i)(k)}\right] \begin{bmatrix} \mathcal{K}(\mathbf{x}_j, \mathbf{z}_1) \\ \vdots \\ \mathcal{K}(\mathbf{x}_j, \mathbf{z}_k) \end{bmatrix} \\
Cov\left[\mathbf{y}_j, \mathbf{a}_i \middle| \mathbf{X}, \lambda, \Sigma\right] &= \left[\mathcal{K}(\mathbf{z}_1, \mathbf{x}_j) \cdots \mathcal{K}(\mathbf{x}_k, z_j)\right] \begin{bmatrix} \mathbf{\Gamma}^{+}_{(1)(i)} \\ \vdots \\ \mathbf{\Gamma}^{+}_{(k)(i)} \end{bmatrix} \\
Cov(\mathbf{a}_i, \mathbf{a}_j | \Sigma,\, \mathbf{X},\, \lambda) &= b\mathbf{\Gamma}^{+}_{(i)(j)}.
\end{aligned}$$

Then

$$\left[\mathrm{vec}\left(\mathcal{A}, \mathbf{Y}\right) | \Sigma,\, \mathbf{X},\, \lambda\right] \sim N_{(k+n)\times d_2}\left[\mathbf{0}, b\left(\begin{array}{c|c} \mathbf{\Gamma}^{+} & \mathbf{\Gamma}^{+}\mathbf{K}_{zx} \\ \hline \mathbf{K}_{xz}\mathbf{\Gamma}^{+} & \mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\left(I_n \otimes \Sigma\right) \end{array}\right)\right].$$

Using the properties of multivariate normal distributions we obtain the conditional distribution $\left[\mathcal{A} | \Sigma,\, \mathbf{X},\, \lambda\right] \sim N_{kd_2}\left(\boldsymbol{\mu}_{\mathbf{Y}}, \Sigma_{\mathbf{Y}}\right)$.

What is left to do now, is to find the expression for the full conditional of Σ. For this task we need that $\mathbf{K}_{zz}\left(I_n \otimes \Sigma^{-1}\right) = \left(I_n \otimes \Sigma^{-1}\right)\mathbf{K}_{zz}$. For any $\{\mathbf{a}_i\}_{i=1}^k, \{\mathbf{b}_i\}_{i=1}^k \subset \mathbb{R}^{d_2}$ we have:

$$\begin{aligned}\left\langle \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i,\cdot)\mathbf{a}_i, \sum_{j=1}^k \mathcal{K}(\mathbf{z}_j,\cdot)\mathbf{b}_j \right\rangle_{\mathcal{H}_\mathcal{K}} &= \sum_{i=1}^k\sum_{j=1}^k \langle \mathcal{K}(\mathbf{z}_i,\cdot)\mathbf{a}_i, \mathcal{K}(\mathbf{z}_j,\cdot)\mathbf{b}_j\rangle_{\mathcal{H}_\mathcal{K}} \\ &= \sum_{i,j=1}^k \langle \mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\mathbf{a}_i, \mathbf{b}_j\rangle_{\mathbb{R}^{d_2}} \qquad (5.11)\\ &= \sum_{i,j=1}^k \mathbf{a}_i^\intercal \mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\Sigma^{-1}\mathbf{b}_j, \qquad (5.12)\end{aligned}$$

equality (5.11) is due to the reproducing property. In a similar way but using the symmetry property of the inner products we obtain

$$\left\langle \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i,\cdot)\mathbf{a}_i, \sum_{j=1}^k \mathcal{K}(\mathbf{z}_j,\cdot)\mathbf{b}_j \right\rangle_{\mathcal{H}_\mathcal{K}} = \sum_{i,j=1}^k \mathbf{a}_i^\intercal \Sigma^{-1}\mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\mathbf{b}_j. \qquad (5.13)$$

Expressions (5.12) and (5.13) must be equal because they are both equal to the same quantity. Furthermore:

$$\sum_{i,j=1}^k \mathbf{a}_i^\intercal \mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\Sigma^{-1}\mathbf{b}_j = \sum_{i,j=1}^k \mathbf{a}_i^\intercal \Sigma^{-1}\mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\mathbf{b}_j$$

$$\text{then} \qquad \mathcal{A}^\intercal \mathbf{K}_{zz}\left(I_k \otimes \Sigma^{-1}\right)\mathcal{B} = \mathcal{A}^\intercal\left(I_k \otimes \Sigma^{-1}\right)\mathbf{K}_{zz}\mathcal{B}$$

In particular, if we take $\mathcal{A} \in \mathcal{M}_{kd_2\times 1}(\mathbb{R})$ the vector column with zeros in all entries but 1 in entry i, and $\mathcal{B} \in \mathcal{M}_{kd_2\times 1}(\mathbb{R})$ the vector column with zeros in all entries but 1 in entry j, we obtain $[\mathbf{K}_{zz}\left(I_k \otimes \Sigma^{-1}\right)]_{i,j} = [\left(I_k \otimes \Sigma^{-1}\right)\mathbf{K}_{zz}]_{i,j}$, hence

$$\mathbf{K}_{zz}\left(I_k \otimes \Sigma^{-1}\right) = \left(I_k \otimes \Sigma^{-1}\right)\mathbf{K}_{zz}. \qquad (5.14)$$

Now we compute the full conditional posterior $[\Sigma|\mathbf{Y},\mathcal{A},\mathbf{X},\lambda]$ writing the log likelihood and priors. In the following equations, C represents a constant that may not be the same in different lines of the equations. The explicit value of C may be obtained by completing the terms required for the expression to be complete.

$$\begin{aligned}-\log\{[\Sigma|\mathbf{Y},\mathcal{A},\mathbf{X},\lambda]\} &= -\log\{[\mathbf{Y}|\mathcal{A},\,\mathbf{X},\,\lambda,\,\Sigma]\times[\mathcal{A}|\Sigma,\,\mathbf{X},\,\lambda]\times[\Sigma]\} + C \\ &= -\frac{1}{2}\log|\Sigma| + \frac{1}{2}\sum_{i=1}^n \left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)^\intercal \Sigma^{-1}\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2}\log\left|b\boldsymbol{\Gamma}^{+}\right|_{+} - \frac{1}{2b}\mathcal{A}^{\intercal}\boldsymbol{\Gamma}\mathcal{A} \\
&- \frac{\nu+d_2+1}{2}\log|\Sigma| + \frac{1}{2}Tr\left\{\mathbf{A}\Sigma^{-1}\right\} + C \\
=& - \frac{1}{2}\left[\log|\Sigma| + \log|\boldsymbol{\Gamma}|_{+}^{-1} + (\nu+d_2+1)\log|\Sigma|\right] \\
&- \frac{1}{2}\left[\sum_{i=1}^{n} Tr\left\{\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)^{\intercal}\Sigma^{-1}\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)\right\} + b^{-1}\ Tr\left\{\mathcal{A}^{\intercal}\boldsymbol{\Gamma}\mathcal{A}\right\}\right] \\
&+ \frac{1}{2}Tr\left\{\mathbf{A}\Sigma^{-1}\right\} + C \\
=& - \frac{1}{2}\left[\log|\Sigma| + \log|\boldsymbol{\Gamma}|_{+}^{-1} + (\nu+d_2+1)\log|\Sigma|\right] \\
&- \frac{1}{2}\left[\sum_{i=1}^{n} Tr\left\{\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)^{\intercal}\Sigma^{-1}\right\} + \frac{1}{b}Tr\left\{\mathcal{A}\mathcal{A}^{\intercal}\boldsymbol{\Gamma}\right\}\right] \\
&+ \frac{1}{2}Tr\left\{\mathbf{A}\Sigma^{-1}\right\} + C \\
=& - \frac{1}{2}\left[\log|\Sigma| + \log|2\mathbf{K}_{zz}\Psi_{\Sigma,k}|_{+}^{-1} + (\nu+d_2+1)\log|\Sigma|\right] \\
&- \frac{1}{2}\left[\sum_{i=1}^{n} Tr\left\{\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)^{\intercal}\Sigma^{-1}\right\}\right] \\
&- \frac{1}{2}\left[\frac{2}{b}\ Tr\left\{\mathcal{A}\mathcal{A}^{\intercal}\mathbf{K}_{zz}\Psi_{\Sigma,k}\right\} + Tr\left\{\mathbf{A}\Sigma^{-1}\right\}\right] + C \qquad (5.15) \\
=& - \frac{1}{2}\left[\log|\Sigma| + \log|\Sigma|^{k} + (\nu+d_2+1)\log|\Sigma|\right] \\
&- \frac{1}{2}\left[\sum_{i=1}^{n} Tr\left\{\left\{\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)^{\intercal} + \mathbf{A}\right\}\Sigma^{-1}\right\}\right] \\
&- \frac{1}{2}\left[n\lambda\ Tr\left\{\sum_{i=1}^{k}\left[\mathcal{A}\mathcal{A}^{\intercal}\mathbf{K}_{zz}\right]_{(i)(i)}\Sigma^{-1}\right\}\right] + C \\
=& - \frac{1}{2}\left[(\nu+1+k) + d_2 + 1\right]\log|\Sigma| \\
&- \frac{1}{2}\left[\sum_{i=1}^{n} Tr\left\{\left\{\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)\left(\mathbf{Y}_i - \eta(\mathbf{x}_i)\right)^{\intercal} + n\lambda\sum_{i=1}^{k}\left[\mathcal{A}\mathcal{A}^{\intercal}\mathbf{K}_{zz}\right]_{(i)(i)} + \mathbf{A}\right\}\Sigma^{-1}\right\}\right],
\end{aligned}
$$

where equation (5.15) is obtained using (5.13). The last expression is the negative of the *log-likelihood* of an inverse Wishart distribution with respective parameters described by:

$$\Sigma|\mathbf{Y}, \mathcal{A} \sim Inv - Wishart\left(\Sigma_p, \nu + k + 1\right).$$

■

One has to observe the interpretation of the deterministic function defined by $\mu_{\mathbf{Y}}$, the mean of the full conditional posterior of $\mathcal{A}$. By Proposition 79, $\mu_{\mathbf{Y}}$ arranged as $\mu_{\mathbf{Y}} = \text{vec}\left(\hat{\mathbf{a}}_1 \cdots \hat{\mathbf{a}}_k\right)$,

$\hat{\mathbf{a}}_i \in \mathbb{R}^{d_2}$ satisfies equation (5.7), then by the arguments in Section 5.1, the function

$$\hat{\eta}(\mathbf{x}) = \sum_{i=1}^{k} \mathcal{K}(\mathbf{z}_i, \mathbf{x})\hat{\mathbf{a}}_i \tag{5.16}$$

solves the minimization problem (5.1) in $\mathcal{H}_{\mathcal{K}}^{\star}$ and as $k \to n$, it solves (5.1) in $\mathcal{H}_{\mathcal{K}}$. This is the interpretation to the proposed regression Bayes model.

Conditional to Σ, point estimates for η in the sampling points $\{\mathbf{x}_i\}_{i=1}^{n}$ can be computed using (5.16) which in matrix notation is written as:

$$\begin{aligned}
\text{vec}\left(\hat{\eta}(\mathbf{x}_1)\cdots\hat{\eta}(\mathbf{x}_n)\right) &= \mathbf{K}_{xz}\mu_Y \\
&= \mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}\mathbf{Y} \\
&= A(\lambda)\mathbf{Y},
\end{aligned}$$

where

$$A(\lambda) = \mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}$$

is the version of expression (2.21) and (2.22) for the current setting. $A(\lambda)$ is in fact a projection matrix but we do not need to prove this statement as this property will not be needed. $A(\lambda)$ will be useful in the Sections ahead.

Nevertheless, the proposed estimator of the function η is the process defined by (5.5) and the marginal posterior $[\mathcal{A}|\mathbf{Y}, \lambda, \mathbf{X}]$. Thus, the estimator for $\eta(\mathbf{x}_i)$ is the marginal posterior mean $[\eta(\mathbf{x}_i)|\mathbf{Y}, \lambda, \mathbf{X}]$ and a predictor for $\eta(\chi_i)$, new value $\chi \in \mathbb{R}^{d_1}$ is the marginal posterior predictive mean $[\eta(\chi_i)|\mathbf{Y}, \lambda, \mathbf{X}]$.

The smoothing parameter $\lambda > 0$ was assumed given. In the next section we propose three methods to choose these parameters conditionally on Σ.

5.2.2 Smoothing parameter selection

The function obtained by approximating the solution to (5.1) through the minimization of (5.6) depend on a known λ, smoothing parameter. The importance of choosing adequate values for λ is that they control the trade off between smoothness of the solution to (5.1) as measured by $\|\cdot\|_{\mathcal{H}}^2$, and how well the solution describe the data as measured by the quadratic

loss function $\sum_{i=1}^{n}\|\eta(\mathbf{x}_i)-\mathbf{y}_i\|_{\mathcal{H}}^2$. The topic of selecting adequate smoothing parameter will be addressed now. Furthermore, we discuss the possibility of choosing bandwidth matrices for improvement of the fitted regression.

5.2.2.1 Smoothing parameter using a unbiased estimate of relative loss method

In this section we conjecture on a generalization of the UERL method to choose the smoothing parameters (Mallows (1973) and Section 2.3.1) for the problem (5.1) with $\mathcal{H}\in\{\eta:\mathbf{R}^{d_1}\to\mathbf{R}^{d_2}\}$. The proposal is not fully proven to have the property we claim, therefore we let the result as a conjecture. Nevertheless, this method seems to provide adequate bandwidth parameters in practice.

In the context of problem (5.1), given $\lambda>0$ and possible extra bandwidth parameters $\{h_i\}_{i=1}^{p}\subset\mathbb{R}^+$ hidden in $\|\cdot\|_{\mathcal{K}_\mathcal{K}}$, Theorem 82 provides conditions for the unique existence of a solution. The performance of $\eta_{\lambda,\theta_1,\cdots,\theta_p}$ solution to the minimizations problem (η_λ by simplicity), can be assessed via the loss function $L(\lambda,\theta_1,\cdots,\theta_p)=L(\lambda)$ with $\Sigma\in\mathcal{M}_{d_2\times d_2}(\mathbb{R})$ positive definite, and defined as

$$L(\lambda)=n^{-1}\sum_{i=1}^{n}\left(\eta_\lambda(\mathbf{x}_i)-\eta(\mathbf{x}_i)\right)^\intercal\Sigma^{-1}\left(\eta_\lambda(\mathbf{x}_i)-\eta(\mathbf{x}_i)\right) \tag{5.17}$$

which is a version of (2.54) for vector values functions.

Remark 19

Model (5.8) leads to an interpretation of having $x^\intercal\Sigma^{-1}x$ as inner product in $\mathbb{R}^{d_2}$. This is one of the reason for the form of the loss function (5.17). Another reason is that we will use these results for regression problems when the response is multivariate normal distributed; the log likelihood of this model has the form of (5.17) so we can apply it later.

Observe that (5.17) as function of $\eta_{\mathbf{X}}=\operatorname{vec}\left(\eta_\lambda(\mathbf{x}_1)\cdots\eta_\lambda(x_n)\right)$ is a convex function, η_λ is unique for each λ (and h_i's), therefore there is a unique minimum for L. The disadvantage of the loss function L is that, since we do not know η, it can not be computed explicitly.

We conjecture that the following score U, a generalization of the score (3.11), estimates L.

$$U(\lambda)=\frac{1}{n}\mathbf{Y}^\intercal\left(I-A(\lambda)\right)^\intercal\left(I_n\otimes\Sigma^{-1}\right)\left(I-A(\lambda)\right)\mathbf{Y}+\frac{2}{n}trA(\lambda). \tag{5.18}$$

Conjecture 20

If $\lim_{\substack{\lambda\to 0\\ n\to\infty}} n\mathbb{E}\left(L(\lambda)\right) = \infty$, *the errors* ϵ_i*'s are independent with common covariance matrix and uniformly bounded fourth moments, then* (5.18) *is an estimate of the relative loss* $L + n^{-1}\epsilon^\intercal\left(I_n \otimes \mathbf{\Sigma}^{-1}\right)\epsilon$ *in the sense*

$$U(\lambda) - L(\lambda) - n^{-1}\epsilon^\intercal\left(I_n \otimes \mathbf{\Sigma}^{-1}\right)\epsilon = o_p(L(\lambda)).$$

We now provide an heuristic proof of the conjecture:

Proof.

Lets define $\eta_{\mathbf{X}} = \mathrm{vec}\left(\eta(\mathbf{x}_1)^\intercal \cdots \eta(\mathbf{x}_n)^\intercal\right)$ and $\Psi_{\Sigma,n} = \left(I_n \otimes \Sigma^{-1}\right)$. nL can be expanded as:

$$\begin{aligned}
nL(\lambda) =& \sum_{i=1}^{n}\left(\eta_\lambda(\mathbf{x}_i) - \eta(\mathbf{x}_i)\right)^\intercal \mathbf{\Sigma}^{-1}\left(\eta_\lambda(\mathbf{x}_i) - \eta(\mathbf{x}_i)\right)\\
=& \sum_{i=1}^{n}\left([A(\lambda)\mathbf{Y}]_{(i)} - \eta(\mathbf{x}_i)\right)^\intercal \mathbf{\Sigma}^{-1}\left([A(\lambda)\mathbf{Y}]_{(i)} - \eta(\mathbf{x}_i)\right)\\
=& \left\{A(\lambda)\mathbf{Y} - \eta_{\mathbf{X}}\right\}^\intercal \Psi_{\Sigma,n}\left\{A(\lambda)\mathbf{Y} - \eta_{\mathbf{X}}\right\}\\
=& \mathbf{Y}^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n} A(\lambda)\mathbf{Y} + \eta_{\mathbf{X}}^\intercal \Psi_{\Sigma,n}\eta_{\mathbf{X}} - 2\mathbf{Y}^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n}\eta_{\mathbf{X}}\\
=& \eta_{\mathbf{X}}^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n} A(\lambda)\eta_{\mathbf{X}} + 2\epsilon^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n} A(\lambda)\eta_{\mathbf{X}} + \epsilon^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n} A(\lambda)\epsilon\\
& - 2\left[\eta_{\mathbf{X}}^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n}\eta_{\mathbf{X}} + \epsilon^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n}\eta_{\mathbf{X}}\right]\\
& + \eta_{\mathbf{X}}^\intercal \Psi_{\Sigma,n}\eta_{\mathbf{X}}\\
=& \eta_{\mathbf{X}}^\intercal\left[I - A(\lambda)\right]^\intercal \Psi_{\Sigma,n}\left[I - A(\lambda)\right]\eta_{\mathbf{X}} - 2\epsilon^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n}\left[I - A(\lambda)\right]\eta_{\mathbf{X}}\\
& + \epsilon^\intercal A(\lambda)^\intercal \Psi_{\Sigma,n} A(\lambda)\epsilon.
\end{aligned}$$

On the other side, U can be rewritten as:

$$\begin{aligned}
n\,U(\lambda) =& \mathbf{Y}^\intercal\left(I - A(\lambda)\right)^\intercal \Psi_{\Sigma,n}\left(I - A(\lambda)\right)\mathbf{Y} + 2trA(\lambda)\\
=& \eta_{\mathbf{X}}^\intercal\left(I - A(\lambda)\right)^\intercal \Psi_{\Sigma,n}\left(I - A(\lambda)\right)\eta_{\mathbf{X}} + 2trA(\lambda)\\
& + \epsilon^\intercal\left(I - A(\lambda)\right)^\intercal \Psi_{\Sigma,n}\left(I - A(\lambda)\right)\epsilon\\
& + 2\epsilon^\intercal\left(I - A(\lambda)\right)^\intercal \Psi_{\Sigma,n}\left(I - A(\lambda)\right)\eta_{\mathbf{X}}\\
=& \eta_{\mathbf{X}}^\intercal\left(I - A(\lambda)\right)^\intercal \Psi_{\Sigma,n}\left(I - A(\lambda)\right)\eta_{\mathbf{X}} + 2trA(\lambda)\\
& + \epsilon^\intercal \Psi_{\Sigma,n}\epsilon + \epsilon^\intercal A(\lambda)\Psi_{\Sigma,n}A(\lambda)\epsilon - \epsilon^\intercal\left[\Psi_{\Sigma,n}A(\lambda) + A(\lambda)\Psi_{\Sigma,n}\right]\epsilon
\end{aligned}$$

$$+ 2\eta_{\mathbf{X}}^{\intercal} \left(I - A(\lambda)\right)^{\intercal} \Psi_{\Sigma,n} \left(I - A(\lambda)\right) \epsilon.$$

Therefore, we can compute with the previous expression of L and U the next equality:

$$\begin{aligned} n\left[U(\lambda) - L(\lambda) - \epsilon^{\intercal}\Psi_{\Sigma,n}\epsilon\right] =& 2trA(\lambda) - \epsilon^{\intercal}\left[\Psi_{\Sigma,n}A(\lambda) + A(\lambda)^{\intercal}\Psi_{\Sigma,n}\right]\epsilon \\ &+ 2\eta_{\mathbf{X}}^{\intercal}\left(I - A(\lambda)\right)^{\intercal}\Psi_{\Sigma,n}\left(I - A(\lambda)\right)\epsilon \\ &+ 2\epsilon^{\intercal}A(\lambda)^{\intercal}\Psi_{\Sigma,n}\left(I - A(\lambda)\right)\eta_{\mathbf{X}} \\ =& 2trA(\lambda) - \epsilon^{\intercal}\left[\Psi_{\Sigma,n}A(\lambda) + A(\lambda)^{\intercal}\Psi_{\Sigma,n}\right]\epsilon \\ &+ 2\eta_{\mathbf{X}}^{\intercal}\left(I - A(\lambda)\right)^{\intercal}\Psi_{\Sigma,n}\epsilon. \end{aligned} \tag{5.19}$$

Observe as well that

$$\begin{aligned} \mathbb{E}\left(\epsilon^{\intercal}\left[\Psi_{\Sigma,n}A(\lambda) + A(\lambda)^{\intercal}\Psi_{\Sigma,n}\right]\epsilon\right) =& tr\left(\left[\Psi_{\Sigma,n}A(\lambda) + A(\lambda)^{\intercal}\Psi_{\Sigma,n}\right]Cov(\epsilon)\right) \\ =& tr\left(\left[\Psi_{\Sigma,n}A(\lambda) + A(\lambda)^{\intercal}\Psi_{\Sigma,n}\right]\Psi_{\Sigma^{-1},n}\right) \\ =& tr\left[\Psi_{\Sigma,n}A(\lambda)\Psi_{\Sigma^{-1},n}\right] + tr\left[A(\lambda)^{\intercal}\Psi_{\Sigma,n}\Psi_{\Sigma^{-1},n}\right] \\ =& tr\left(\left[\Psi_{\Sigma,n}A(\lambda) + A(\lambda)^{\intercal}\Psi_{\Sigma,n}\right]\Psi_{\Sigma^{-1},n}\right) \\ =& tr\left[\Psi_{\Sigma,n}A(\lambda)\Psi_{\Sigma^{-1},n}\right] + tr\left[A(\lambda)^{\intercal}\Psi_{\Sigma,n}\Psi_{\Sigma^{-1},n}\right] \\ =& tr\left[\Psi_{\Sigma^{-1},n}\Psi_{\Sigma,n}A(\lambda)\right] + tr\left[A(\lambda)^{\intercal}\right] \\ =& 2trA(\lambda). \end{aligned} \tag{5.20}$$

We think the conjecture follows by using (5.19), (5.20) and the next expressions that would need to be proven as well using the eigenvalue decomposition $A(\lambda) = PDP^{\intercal}$, that the eigenvalues are between 0 and 1, common covariance matrix and uniformly fourth bounded moments:

$$\begin{aligned} L(\lambda) - \mathbb{E}L(\lambda) &= o_p\left(\mathbb{E}L(\lambda)\right) \\ \frac{1}{n}\eta_{\mathbf{X}}^{\intercal}(I - A(\lambda))\Psi_{\Sigma^{-1},n}\epsilon &= o_p\left(\mathbb{E}L(\lambda)\right) \\ \frac{1}{n}\left\{2trA(\lambda) - \epsilon^{\intercal}\left[\Psi_{\Sigma,n}A(\lambda) + A(\lambda)\Psi_{\Sigma,n}\right]\epsilon\right\} &= o_p\left(\mathbb{E}L(\lambda)\right). \end{aligned}$$

The heuristic proof ends here. But a strict mathematical justification would need to take into account that the minimizers of U and L are stochastic. ■

Since $\frac{1}{n}\epsilon^\intercal \Psi_{\Sigma,n}\epsilon$ does not depend on λ, and if the conclusion of Conjecture (20) holds, then $U(\lambda)$ tracks $L(\lambda)$ closely. By consistency with the previous chapters we call the minimizer λ_u of (5.18) the *Unbiased Estimate of Relative Loss* (UERL) estimate of λ.

5.2.2.2 Smoothing parameter using a generalized cross validation type method

In Section 2.3.2, a cross validation procedure to select bandwidth parameters for the penalized least squares minimization problem (1.1) was reviewed. We now attempt to generalize such method and adjust the ideas to select adequate bandwidth parameters for the minimization problem (5.1) in the space of vector valued multivariate functions in the euclidean space. We propose a generalization to the generalized cross validation method (2.59) and conjecture about the properties of our proposed score function.

For the observed realizations $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$, we minimize, with respect to λ, the score V_0 described by the expression:

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \left(\eta_\lambda^{[i]}(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal \boldsymbol{\Sigma}^{-1} \left(\eta_\lambda^{[i]}(\mathbf{x}_i) - \mathbf{y}_i\right),$$

where $\eta_\lambda^{[k]}$ is the minimizer of the functional

$$\sum_{\substack{i=1 \\ i\neq k}}^{n} (\mathbf{y}_i - \eta(\mathbf{x}_i))^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \eta(\mathbf{x}_i)) + n\lambda J(\eta). \tag{5.21}$$

It is not necessary to solve (5.21) n times but the delete-one operation can be done analytically as the following Lemma states.

Lemma 21

The minimizer $\eta_\lambda^{[k]}$ of the delete-one functional (5.21) minimizes the full data functional

$$(\tilde{\mathbf{y}}_k - \eta(\mathbf{x}_k))^\intercal \Sigma^{-1} (\tilde{\mathbf{y}}_k - \eta(\mathbf{x}_k)) + \sum_{\substack{i=1 \\ i\neq k}}^{n} (\mathbf{y}_i - \eta(\mathbf{x}_i))^\intercal \Sigma^{-1} (\mathbf{y}_i - \eta(\mathbf{x}_i)) + n\lambda J(\eta),$$

where $\tilde{\mathbf{y}}_k = \eta_\lambda^{[k]}(\mathbf{x}_k)$.

Proof.

Observe that for all $\eta \neq \eta_\lambda^{[k]}$

$$\left(\tilde{\mathbf{y}}_k - \eta_\lambda^{[k]}(\mathbf{x}_k)\right)^\intercal \Sigma^{-1} \left(\tilde{\mathbf{y}}_k - \eta_\lambda^{[k]}(\mathbf{x}_k)\right) + \sum_{\substack{i=1 \\ i\neq k}}^{n} \left(\mathbf{y}_i - \eta_\lambda^{[k]}(\mathbf{x}_i)\right)^\intercal \Sigma^{-1} \left(\mathbf{y}_i - \eta_\lambda^{[k]}(\mathbf{x}_i)\right) + n\lambda J(\eta_\lambda^{[k]})$$

$$
\begin{aligned}
&= \sum_{\substack{i=1 \\ i \neq k}}^{n} \left(\mathbf{y}_i - \eta_\lambda^{[k]}(\mathbf{x}_i)\right)^\intercal \Sigma^{-1} \left(\mathbf{y}_i - \eta_\lambda^{[k]}(\mathbf{x}_i)\right) + n\lambda J(\eta) \\
&< \sum_{\substack{i=1 \\ i \neq k}}^{n} \left(\mathbf{y}_i - \eta(\mathbf{x}_i)\right)^\intercal \mathbf{\Sigma}^{-1} \left(\mathbf{y}_i - \eta(\mathbf{x}_i)\right) + n\lambda J(\eta) \\
&\leq \left(\tilde{\mathbf{y}}_k - \eta(\mathbf{x}_k)\right)^\intercal \Sigma^{-1} \left(\tilde{\mathbf{y}}_k - \eta(\mathbf{x}_k)\right) + \sum_{\substack{i=1 \\ i \neq k}}^{n} \left(\mathbf{y}_i - \eta(\mathbf{x}_i)\right)^\intercal \mathbf{\Sigma}^{-1} \left(\mathbf{y}_i - \eta(\mathbf{x}_i)\right) + n\lambda J(\eta)
\end{aligned}
$$

■

We aim to rewrite V_0 using Lemma 21 in a way that its computation is more efficient. First we have:

$$
\begin{aligned}
&\text{by Lemma 21} && \operatorname{vec}\left(\eta_\lambda^{[k]}(\mathbf{x}_1) \cdots \eta_\lambda^{[k]}(\mathbf{x}_n)\right) = A(\lambda) \operatorname{vec}\left(\mathbf{y}_1 \cdots \mathbf{y}_{k-1}\, \eta_\lambda^{[k]}(\mathbf{x}_k)\, \mathbf{y}_{k+1} \cdots \mathbf{y}_n\right), \\
&\text{then} && \operatorname{vec}\left(\eta_\lambda(x_1) - \eta_\lambda^{[k]}(\mathbf{x}_1) \cdots \eta_\lambda(x_n) - \eta_\lambda^{[k]}(\mathbf{x}_n)\right) = A(\lambda) \operatorname{vec}\left(\mathbf{0}_{d_2} \cdots \mathbf{0}_{d_2}\, \mathbf{y}_k - \eta_\lambda^{[k]}(\mathbf{x}_k) \cdots \mathbf{0}_{d_2}\right), \\
&\text{then} && \eta_\lambda(\mathbf{x}_j) - \eta_\lambda^{[k]}(\mathbf{x}_j) = [A(\lambda)]_{(j)(k)} \left(\mathbf{y}_k - \eta_\lambda^{[k]}(\mathbf{x}_k)\right), \\
&\text{then} && \eta_\lambda(\mathbf{x}_i) - [A(\lambda)]_{(i)(i)} \mathbf{y}_i = \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right) \eta_\lambda^{[i]}(\mathbf{x}_i), \\
&\text{then} && \eta_\lambda(\mathbf{x}_i) - [A(\lambda)]_{(i)(i)} \mathbf{y}_i - \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right) \mathbf{y}_i = \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right) \left(\eta_\lambda^{[i]}(\mathbf{x}_i) - \mathbf{y}_i\right), \\
&\text{then} && \eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i = \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right) \left(\eta_\lambda^{[i]}(\mathbf{x}_i) - \mathbf{y}_i\right), \\
&\text{then} && \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)^{-1} \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right) = \left(\eta_\lambda^{[i]}(\mathbf{x}_i) - \mathbf{y}_i\right), \\
&\text{then} && \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)^{-1} \Sigma^{-1} \ldots \\
& && \ldots \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)^{-1} \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right) = \left(\eta_\lambda^{[i]}(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal \Sigma^{-1} \left(\eta_\lambda^{[i]}(\mathbf{x}_i) - \mathbf{y}_i\right),
\end{aligned}
$$

where $[A(\lambda)]_{(i)(j)}$ is the $(i,j)th$ $d_2 \times d_2$ sub-matrix of $A(\lambda)$. Therefore

$$
\begin{aligned}
V_0(\lambda) &= \sum_{i=1}^{n} \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal \left[\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right) \Sigma \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)\right]^{-1} \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right) \\
&= \sum_{i=1}^{n} tr\left\{\left[\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right) \Sigma \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)\right]^{-1} \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right) \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal\right\}.
\end{aligned}
$$

Extrapolating the ideas from the uni-variate case, Section 2.3.2, not all sampling points contribute equally to the estimation of η. We could use the next weighted version of $V_0(\lambda)$:

$$
V_1(\lambda) = \sum_{i=1}^{n} tr\left\{\omega_i \left[\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right) \Sigma \left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)\right]^{-1} \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right) \left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal\right\}.
$$

If $\omega_i = \frac{1}{n}\left\{\sum_{i=1}^{n}\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)\Sigma\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)\right\}^{-1}\left[\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)\Sigma\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)\right]$ is chosen, a generalized cross validation score is obtained. Such score is an attempt to extend the method proposed by Wahba and Craven (1978), Li (1986), Gu (2013), Section 2.3.2. Let

$$\begin{aligned}
V_2(\lambda) &= \sum_{i=1}^{n} tr\left\{\left\{\sum_{j=1}^{n}\left(I_{d_2} - [A(\lambda)]_{(j)(j)}\right)\Sigma\left(I_{d_2} - [A(\lambda)]_{(j)(j)}\right)\right\}^{-1}\left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right)\left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal\right\} \\
&= \sum_{i=1}^{n}\left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal\left\{\sum_{j=1}^{n}\left(I_{d_2} - [A(\lambda)]_{(j)(j)}\right)\Sigma\left(I_{d_2} - [A(\lambda)]_{(j)(j)}\right)\right\}^{-1}\left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right) \\
&= \sum_{i=1}^{n}\left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right)^\intercal \mathbf{F}_\lambda^{-1}\left(\eta_\lambda(\mathbf{x}_i) - \mathbf{y}_i\right) \\
&= \sum_{i=1}^{n}\left[A(\lambda)\mathbf{Y} - \mathbf{Y}\right]_{(i)}^\intercal \mathbf{F}_\lambda^{-1}\left[A(\lambda)\mathbf{Y} - \mathbf{Y}\right]_{(i)} \\
&= \mathbf{Y}^\intercal\left[I_{nd_2} - A(\lambda)\right]^\intercal \Psi_{\mathbf{F}_\lambda, n}\left[I_{nd_2} - A(\lambda)\right]\mathbf{Y},
\end{aligned}$$

where $\mathbf{F}_\lambda = \sum_{i=1}^{n}\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)\Sigma\left(I_{d_2} - [A(\lambda)]_{(i)(i)}\right)$. Again, Conditions 3 and 4 provide enough assumptions so that (2.58), the version of V_2 for the real nonparametric problem regression problem, is a consistent estimator of the relative loss. We were unable to provide conditions so that $\mathcal{V}_2$ is a consistent estimator of the relative loss (5.17) and we leave it as an open problem.

Problem 22 *What are minimal conditions so that V_2 is a consistent estimator of the loss function (5.17)?.*

A conjecture to solve this problem is the following.

Conjecture 23 *If* $\lim_{\substack{\lambda\to 0\\ n\to\infty}} n\mathbb{E}\left(L(\lambda)\right) = \infty$. $\lim_{\substack{\lambda\to 0\\ n\to\infty}} \frac{\{n^{-1}trA(\lambda)\}^2}{n^{-1}tr\left[A(\lambda)\Psi_{\Sigma,n}A(\lambda)\Psi_{\Sigma^{-1},n}\right]} = 0$, *the errors ϵ_i's are independent with common covariance matrix and uniformly bounded fourth moments, then V_2 is an estimate of the relative loss (5.17) in the sense*

$$n^{-1}V_2(\lambda) - L(\lambda) - n^{-1}\epsilon^\intercal\left(I_n \otimes \Sigma^{-1}\right)\epsilon = o_p(L(\lambda)).$$

It was mentioned for the univariate response regression problem, the respective version of V_2, that the minimization of the score (2.58) occasionally delivers λ that leads to sever undersmoothed regressions. Kim and Gu (2004) for the real regression problem suggest a modified

version with a fudge factor α, expression (2.59). Since there are not strong reasons to doubt about occasional severe under smooth using our proposal V_2, we propose a version with a fudge factor α as well.

$$V(\lambda, \alpha) = \mathbf{Y}^\intercal \left[I_{nd_2} - A(\lambda)\right]^\intercal \Psi_{\mathbf{F}_\lambda(\alpha),n} \left[I_{nd_2} - A(\lambda)\right] \mathbf{Y}, \tag{5.22}$$

with $\mathbf{F}_\lambda(\alpha) = \sum_{i=1}^n \left(I_{d_2} - \alpha[A(\lambda)]_{(i)(i)}\right) \Sigma \left(I_{d_2} - \alpha[A(\lambda)]_{(i)(i)}\right)$.

It is indeed needed to study the properties (5.22) and answer the basic questions such as: is it a consistent estimate of V in any sense for some minimal conditions? (conjecture 23), what adequate value of α can be taken? For now we decide to use $\alpha = 1.4$ as an empirical value which was recommended by Kim and Gu (2004) in the case of the problem for real valued regression functions. Kim *et. al.* and from Chapter 3 we conclude that the empirical value $\alpha = 1.4$ provides adequate performance over a range of simulation settings. By the same arguments as by Kim *et. al.*, an optimal value α, if indeed there is an optimal one, would depend on the true function η and possibly other factors. Let us denote as λ_v the global minimizer of (5.22).

5.2.2.3 Smoothing parameter selection using a maximum likelihood method under Bayesian model

Using same notation as before, we propose a third score so that its minimizer can be seen as a smoothing parameter to be used in the minimization problem (5.1).

$$\begin{aligned}
\mathsf{M}(\lambda) =& \lambda \mathbf{Y}^\intercal \left\{ \mathbf{K}_{xz} \left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zx} + \frac{n\lambda}{2} \left[\left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zz} + \mathbf{K}_{zz} \left(I_n \otimes \Sigma^{-1}\right) \right] \right\} \mathbf{Y} \\
& \times \left| \frac{2}{n\lambda} \mathbf{K}_{xz} \left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zx} + \left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zz} + \mathbf{K}_{zz} \left(I_n \otimes \Sigma^{-1}\right) \right|_+^{-\frac{1}{nd_2}} \\
\propto & \lambda^{1-\frac{k^\star}{nd_2}} \mathbf{Y}^\intercal \left\{ \mathbf{K}_{xz} \left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zx} + \frac{n\lambda}{2} \left[\left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zz} + \mathbf{K}_{zz} \left(I_n \otimes \Sigma^{-1}\right) \right] \right\} \mathbf{Y} \\
& \times \left| \mathbf{K}_{xz} \left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zx} + \frac{n\lambda}{2} \left[\left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zz} + \mathbf{K}_{zz} \left(I_n \otimes \Sigma^{-1}\right) \right] \right|_+^{\frac{1}{nd_2}}
\end{aligned} \tag{5.23}$$

where $|A|_+$ is the product of the positive eigenvalues of A and

$$k^\star = rank \left\{ \left(I_n \otimes \Sigma^{-1}\right) \mathbf{K}_{zz} + \mathbf{K}_{zz} \left(I_n \otimes \Sigma^{-1}\right) \right\}. \tag{5.24}$$

The function M is a generalization of the score described in Section 2.72, and was obtained following the lines of Section 2.3.3. Such method is not designed to select smoothing parameters

by minimizing any loss function like (5.23) instead, the smoothing parameters arises in the context of some parameters in a Bayesian model based fitting approach. Let us denote as λ_m the global minimizer of M.

In order to visualize the scores described in this section, we simulated a training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{100} \subset \mathbb{R}^2 \times \mathbb{R}^2$ with the form $\mathbf{y}_i = (\eta_1(\mathbf{x}_i), \eta_2(\mathbf{x}_i)) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N_2\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$. η_1 is described by (3.17) and η_2 is described by (5.25).

$$\eta((x_{(1)}, x_{(2)})) = \left[4 + \frac{\sin(\frac{1}{2}\pi x_{(1)})}{1 + 4x_{(1)}^2 \mathbf{1}(x_{(1)} > 0)}\right] + \left[\sin(x_{(2)}) + \cos(x_{(2)}) + x_{(2)}\right]. \tag{5.25}$$

Figure 5.1 next, shows and example of the plotted scores obtained with simulated data, $\mathbf{x}_i \in \mathbb{R}^2$, $\mathbf{y}_i \in \mathbb{R}^2$. For each plot and only for this example we uses four algorithms in order to compute the minimizer, we use the software R functions *GenSA* Yang Xiang et al. (2013), *optim*, *optimize*, R Core Team (2016) and *nlm0* Gu (2014). It is noticeable that the Restrictive Maximum likelihood score is numerically unstable and a smoothed version was used to find the minimum. All scores have its respective global minimum and the three of them are close to each other. The GCV score has an obvious local minimum; one has to be careful that the numerical method used to minimize the score does not return a local minimum.

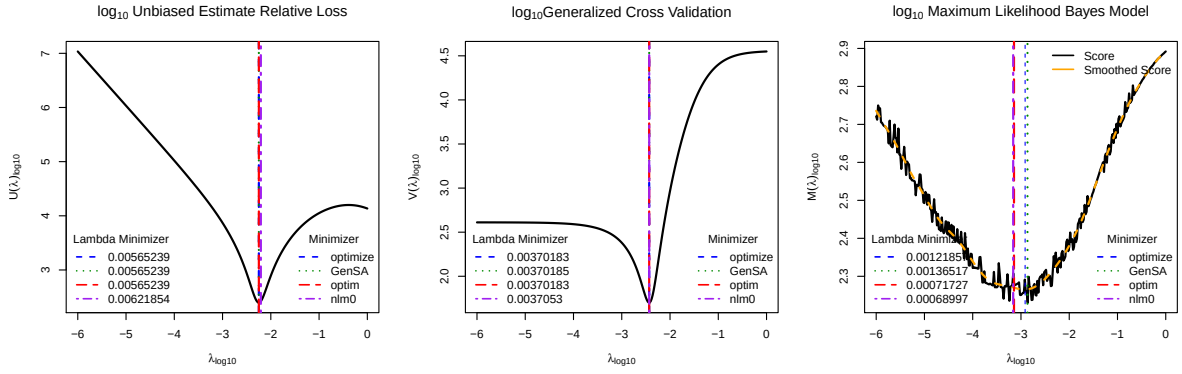

Figure 5.1 Example Score for Single Smoothing Parameter, Vector Multivariate Regression Problem. The global minimum of the scores provide a selection for the smoothing parameter in the minimization problem (5.1). Data simulated with $n = 100$, $\Sigma = \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)$. $\eta = (\eta_1, \eta_2)$ where η_1 and η_2 are described by (3.17) and (5.25) respectively. Plots computed using known Σ, $k = 60$

5.2.3 Possible generalization to diagonal bandwidth matrices and to full bandwidth matrices

Given reproducing kernel $\mathcal{K}$, its associated *RKHS* $\mathcal{H}_{\mathcal{K}}$ and the minimization problem (5.1), we proposed three methods to chose the univariate smoothing parameter $\lambda > 0$ by introducing first, a loss function (5.17) consistent with the inner product assumed for $\mathbb{R}^{d_2}$, second, the proposed scores were discussed to closely follow the form of the loss function, finally, we argued that the minimizer of such scores is an estimator of the minimizer of the loss function (5.17).

We now propose an extension of the methods to select bandwidth matrices for the minimization problem (5.1). We assume we have available a reproducing kernel $\mathcal{K}$ in the space $\mathcal{H}_{\mathcal{K}} \subset \left\{ \eta : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2} \middle| \, \|\eta\|_{\mathcal{H}_{\mathcal{K}}} < \infty \right\}$. Lets define the parametric set of reproducing kernels

$$\mathfrak{S}_{\mathcal{K}} = \left\{ H\mathcal{K}(\cdot,\cdot)H | H \in \mathcal{M}^{++}_{d_2 \times d_2}(\mathbb{R}) \right\}.$$

Each element of $\mathfrak{S}_{\mathcal{K}}$ is a reproducing kernel because we multiplied by both sides with a positive definite linear operator (Belkin et al. (2005); Micchelli and Pontil (2005); Caponnetto et al. (2008)). Each reproducing kernel $H\mathcal{K}H$ has its respective *RKHS* (Micchelli and Pontil (2005)) and we can define the minimization problem (5.1) using the induced norm $\|\cdot\|_{H\mathcal{K}H}$ in such *RKHS* . Then, we can repeat, for fixed $H \in \mathcal{M}^{++}_{d_2 \times d_2}(\mathbb{R})$, all the arguments provided in Sections 5.2.2.1, 5.2.2.2 and 5.2.2.3 within each *RKHS* in $\mathfrak{S}_{\mathcal{K}}$: we take the loss function (5.17) and approximate the minimizer either of the either of the scores U, V or M. Now, such score functions depend as well of a positive definite matrix H trough the projection matrix $A(\lambda, H)$. We can minimize over such scores over $\lambda > 0$ and H in the cone of positive definite matrices. In order to avoid identifiability issues, we need to set a constrain on λ and H, we can put $\lambda = 1$.

If we set $\lambda = 1$ and assume that $H = \begin{pmatrix} H_{11} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & H_{d_2 d_2} \end{pmatrix}$ is diagonal, $H_{ii} > 0$, the minimization of the scores (5.18), (5.22) and (5.23) is carried over the positive quadrant of $\mathbb{R}^{d_2}$. If we set $\lambda = 1$ and H is required to be general positive definite matrix, we would need to minimize over the cone of positive definite matrices. A way to achieve this minimization is by using the Cholesky decomposition $H = LL^{\intercal}$ (Banerjee and Roy (2014)). The Cholesky decomposition is unique for positive definite matrices, where L is a lower triangular matrix, the diagonal has only positive values and the lower off diagonal elements can take on any value. The parametrization

provided by Cholesky decomposition is helpful when using the algorithms to minimize the multivariate function (5.18), (5.22) and (5.23) in the correct constrained space. Furthermore, if we want to use full positive definite matrices H in $\mathfrak{S}_{\mathcal{K}}$, it is expensive to compute the Cholesky decomposition each time the scores are needed to be evaluated, instead we could use kernels of the form $L\mathcal{K}(\cdot,\cdot)L^{\intercal}$ with L a matrix similar to the ones obtained from the Cholesky decomposition of positive definite matrices. Such operators, $L\mathcal{K}(\cdot,\cdot)L^{\intercal}$, are indeed reproducing kernels of a *RKHS* (Caponnetto et al. (2008)).

As visualization of the function scores defined in the way just described, we simulated a training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{100} \subset \mathbb{R}^2 \times \mathbb{R}^2$ with the form $\mathbf{y}_i = (\eta_1(\mathbf{x}_i), \eta_2(\mathbf{x}_i)) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$. η_1 is described by (3.17) and η_2 is described by (5.25). Figure 5.1 shows and example of the plotted scores obtained with the simulated data. For each plot we used the *L-BFGS-B* method *optim* the from R-software function (R Core Team (2016)).

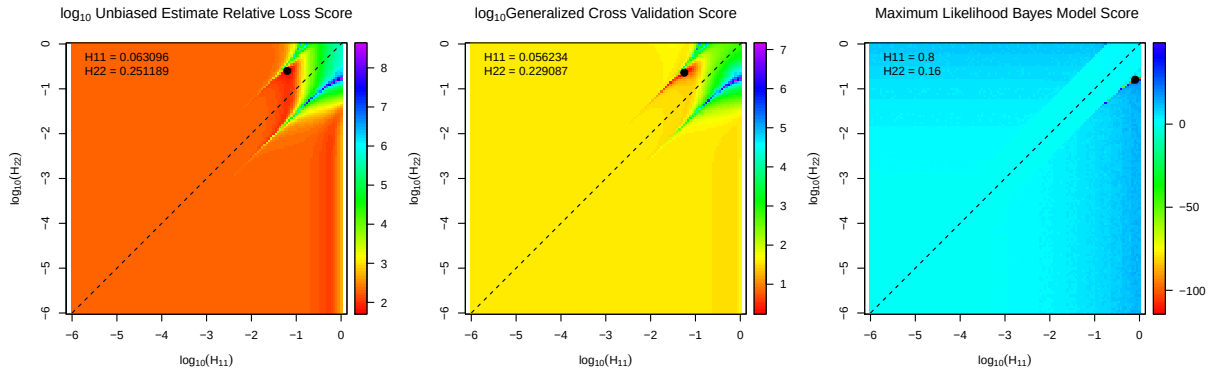

Figure 5.2 Example Scores Functions to Select Diagonal Bandwidth Matrices. Vector Multivariate Regression Problem. Data simulated with $n = 100$, $k = 60$, $\Sigma = \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)$. $\eta = (\eta_1, \eta_2)$ where η_1 is described by (3.17) and η_2 is described by (5.25). The minimizer of the scores are represented in the plots by the dot. It was not possible to minimize the score *Restrictive Maximum Likelihood* because of numerical instability of the evaluation of M, but we visually selected the minimizer.

Remark 24

The main difficulty, we found, of selecting bandwidth matrices by minimizing either U *or* V*, is that there exists local minimums. Depending on the starting point to search for the minimums, one can not ensure to have found the global minimum. The global minimum in Figure* 5.2

left and center plots, were found by providing an adequate starting point to be employed in the search L-BFGS-B algorithm. The minimum indicated in the right graph, Figure 5.2, for M*, was found by looking at the plot.*

There are still open problems regarding to selecting bandwidth matrices and we propose these methods as starting points for further research.

5.2.4 Regression model with selection of smoothing parameters and estimator of observed error covariance

Proposition 18 describes a Bayesian model with the property that the deterministic function $\hat{\eta}(\chi) = \mathbb{E}[\eta(\chi)|\mathbf{Y}, \Sigma, \lambda]$ approximates the minimization of (5.1), and thus, we used the process defined by $[\eta(\cdot)|\mathbf{Y}, \Sigma, \lambda]$ as estimator for the regression function η in (5.8). Thus, as Proposition 18 is now, we need to have selected first the bandwidth parameters $\lambda > 0$ and $H \; in \mathcal{M}^{++}_{d_2 \times d_2}(\mathbb{R})$, which can be done as described in Sections 5.2.2 and 5.2.3. Nevertheless, the score functions U, V and M depend on the covariance matrix Σ. One can provide an external estimator $\hat{\Sigma}_{ext}$, select the bandwidth parameters taking $\Sigma = \hat{\Sigma}_{ext}$ and finally fit the model in Proposition 18 using $\Sigma = \hat{\Sigma}_{ext}$ and the selected smoothing parameters.

Examples of estimators $\hat{\Sigma}$ without the need of estimating η could be obtained using difference covariance estimators, generalizations of expressions (4.11), (4.12), (4.13) or (4.14). We admit these type of estimators for Σ are only conjectures from our side because we did not find previous work on this subject. Nevertheless, we investigated the possible matrix estimate version of (4.12) using the expression

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{y}_{i,1}) (\mathbf{y}_i - \mathbf{y}_{i,1})^{\intercal}, \tag{5.26}$$

where $\mathbf{y}_{i,1}$ is the first nearest neighbor (Definition 44) of $\mathbf{y}_i$. Our small simulations for $\Sigma \in \mathcal{M}_{2\times 2}(\mathbb{R})$ provided positive results in the sense that $\hat{\Sigma}$ seemed to be an unbiased estimator with small variability, $\hat{\Sigma} \approx \Sigma$.

From our experience with simulations in Chapters 3 and 4, we think that a Bayesian approach to estimate η with difference method for $\hat{\Sigma}$ would be surpassed by a model that estimates η, Σ and selects the bandwidth parameters simultaneously. Such model will be our focus of

attention in the following sections. Nevertheless, the former approach is computationally less intensive because the posterior distributions are known and no MCMC methods are needed.

Proposition 25

Let $\mathcal{H}$ a RKHS of functions with domain $\mathbb{R}^{d_1}$ and rank in $\mathbb{R}^{d_2}$. Let $\mathcal{K}$ the reproducing kernel of $\mathcal{H}$. Let the pairs $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ be an observed labeled training set, let $\mathbf{Z} := \{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n =: \mathbf{X}$. Let $\mathbf{Y} = \text{vec}(\mathbf{y}_1\, \mathbf{y}_2 \cdots \mathbf{y}_n)$, let $\mathcal{A} = \text{vec}(\mathbf{a}_1\, \mathbf{a}_2 \cdots \mathbf{a}_k)$ with $\mathbf{a}_i \in \mathbb{R}^{d_2}$, let $\mathbf{K}_{xz} \in \mathcal{M}_{nd_2 \times kd_2}(\mathbb{R})$ be a block matrix with i,jth block $\mathcal{K}(\mathbf{x}_i, \mathbf{z}_j)$, let $\mathbf{K}_{zz} \in \mathcal{M}_{kd_2 \times xd_2}(\mathbb{R})$ be a another block matrix with i,jth block $\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)$, let $\mathbf{K}_{zx} = \mathbf{K}_{xz}^\intercal$, $\Psi_{\Sigma,n} = \left(I_n \otimes \Sigma^{-1}\right)$ and $\mathbf{\Gamma} = \Psi_{\Sigma,k}\mathbf{K}_{zz} + \mathbf{K}_{zz}\Psi_{\Sigma,k}$. Consider the model

$$\begin{aligned}
\mathbf{y}_i &= \eta(\mathbf{x}_i) + \epsilon_i, \\
\eta(\mathbf{x}) &= \sum_{i=1}^{k} \mathcal{K}(\mathbf{z}_i, \mathbf{x})\mathbf{a}_i, \\
\epsilon_i &\overset{iid}{\sim} N_{d_2}(\mathbf{0}, \Sigma),
\end{aligned}$$

with priors on the parameters

$$\begin{aligned}
\mathcal{A}|\Sigma, H, \lambda \sim & N_{kd_2}\left[\mathbf{0}, \frac{2}{n\lambda}\mathbf{\Gamma}^+\right], \\
\mathbf{P}(H < H_0|\Sigma = S, \mathbf{X} = \mathbf{x}) = & \int_{\mathbb{R}^{nd_2}} \mathbf{1}\left\{H_0 < \underset{x=1,\, H_1 \;\; pos.def.}{\arg\min} U(x, H_1|S, \mathbf{y}, \mathbf{x})\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}), \\
\lambda|\Sigma = & 1 \;\; \textit{almost surely} \\
\Sigma \sim & Inv - Wishart(\mathbf{A}, \nu),
\end{aligned}$$

where $\mathbf{\Gamma}^+$ is the Moore-Penrose inverse of the matrix $\mathbf{\Gamma}$. Unless stated otherwise, independence is assumed. The event $H < H_0$ for two positive definite matrices denote that $H_0 - H$ is positive definite.

Alternatively, one can assign the following priors to the smoothing parameters.

$$\mathbf{P}(\lambda \geq \lambda_0 | \Sigma = S, \mathbf{X} = \mathbf{x}) = \int_{\mathbb{R}^{nd_2}} \mathbf{1}\left\{\lambda_0 \geq \underset{x>0, H_1 = I_{d_2}}{\arg\min}\ U(x, H_1 | S, \mathbf{y}, \mathbf{x})\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}), \tag{5.27}$$

$$H|\Sigma = I_{d_2} \text{ almost surely.}$$

Remark 26

If a bandwidth matrix H is required, such as in the former priors for H and λ, the matrices $\mathbf{K}_{zz}$ and $\mathbf{K}_{xz}$ are function of H, thus $\mathbf{\Gamma}$ is function of H as well. If it is assumed that $H = I_{d_2}$ and only $\lambda > 0$ is needed to be selected, as in the later alternative priors (5.27) , then neither $\mathbf{K}_{zz}$ nor $\mathbf{K}_{xz}$ depend of smoothing parameters.

Then, the joint posterior of the parameters exists and the full conditional posteriors are

$$\begin{aligned}
\mathcal{A}|\mathbf{Y}, \mathbf{\Sigma}, \lambda, H \sim & N_{kd_2}\left[\mu_{\mathbf{Y}}, \mathbf{\Sigma}_{\mathbf{Y}}\right], \\
& \mu_{\mathbf{Y}} = \mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}\mathbf{Y}, \\
& \mathbf{\Sigma}_{\mathbf{Y}} = \frac{n\lambda}{2}\left\{\mathbf{\Gamma}^{+} - \mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\right\}, \\
(H|\Sigma, \mathbf{Y}) = & \underset{x=1,\ H\ pos.def.}{\arg\min}\ \{U(x, H|\Sigma, \mathbf{Y})\} \text{ almost surely}, \\
(\lambda|\Sigma, \mathbf{Y}) = & 1 \text{ almost surely}, \\
\Sigma|\mathbf{Y}, \mathcal{A}, \lambda, H \sim & Inv - Wishart\left(\Sigma_p, \nu + k + 1\right), \\
& \Sigma_p = \mathbf{A} + \frac{n\lambda}{2}\sum_{i,j=1}^{k} H\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)H\mathbf{a}_i\mathbf{a}_j^{\intercal} \\
& \quad + \begin{pmatrix} \mathbf{y}_1 - \eta(\mathbf{x}_1) & \cdots & \mathbf{y}_n - \eta(\mathbf{x}_n) \end{pmatrix} \begin{pmatrix} (\mathbf{y}_1 - \eta(\mathbf{x}_1))^{\intercal} \\ \vdots \\ (\mathbf{y}_n - \eta(\mathbf{x}_n))^{\intercal} \end{pmatrix}.
\end{aligned}$$

If the priors (5.27) were chosen, the corresponding full conditional posteriors are

$$\begin{aligned}
H|\Sigma &= I_{d_2} \text{ almost surely}, \\
\lambda|\Sigma &= \underset{x>0, H=I}{\arg\min}\ \{U(x, H|\Sigma, \mathbf{Y})\} \text{ almost surely}.
\end{aligned}$$

Proof.

The joint posterior distribution exists because all the priors are proper. Conditional on λ, H and Σ, the full conditional posterior of $\mathcal{A}$ is already proven to have the claimed form because of Proposition 18. The full conditional posterior of λ or H can be obtained directly observing that the prior was defined with the form $\int f_{\lambda|\mathbf{y}}(\lambda)dF_Y(\mathbf{y})$, thus, the full conditional posterior is given by the distribution defined by $f_{\lambda|\mathbf{y}}$. The full conditional posterior of Σ is obtained as in the proof of Proposition 18. ■

Additionally, the model in Proposition 25 can be modified through the priors on λ and H conditional on Σ by minimizing the scores V or M instead of using U. For the expressions of U, V and M see (5.18), (5.22) and (5.23) respectively.

Remark 27

One have to use caution in the process of fitting the model because numerical problems to find the minimum of the scores U, V *or* M *as function of positive definite matrices* H, *as was described in Remark* 24.

5.2.5 Implementation and example of estimation

In this section we provide details on the fitting procedure of model from Proposition 25. It is needed in Proposition 25 a matrix-valued positive definite kernel $\mathcal{K} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_1} \to \mathcal{M}^{+}_{d_2 \times d_2}(\mathbb{R})$. Such Kernel completely defines the Hilbert space $\mathcal{H}_{\mathcal{K}}$, its inner product and the respective norm $\|\cdot\|_{\mathcal{H}_{\mathcal{K}}}$ (Carmeli et al., 2006, Proposition 2.3). Therefore, the minimization problem (5.1) and its interpretation is completely defined by $\mathcal{K}$. In a sense, we are proceeding to estimate η using a different direction of arguments as we did for real valued regression functions in the corresponding discussions from Chapters 2, 3 and 4. In those Sections, we started with a functional minimization problem (1.1) such as thin plate splines (Section 2.1.1) or tensor thin plate splines (Section 2.1.2). Then, we computed its real reproducing kernel R_J or R_1 using different techniques as described in Sections 2.1.1.1, 2.1.1.2, or expression (2.40); the discussion followed noticing that an approximate solution to the functional minimization problems are of

the form (2.49) which is mostly determined by the reproducing kernel. Now, for the vector valued regression problem, we start with a matrix valued reproducing kernel $\mathcal{K}$, and even, when in principle we do not know explicitly the induced $\|\cdot\|_{\mathcal{H}}$, we are using the solution to the minimization problem (5.1).

A possible generalization of the reproducing kernel for thin plate splines for vector valued functions in $\mathbb{R}^3$ and its respective induced norm $\|\cdot\|_{\mathcal{H}_\mathcal{K}}$ was described in (Benbourhim and Bouhamidi (2005)). The norm in such *RKHS* measures smoothness in terms of derivatives. More kernels for vector valued functions in $\mathbb{R}^2$ or $\mathbb{R}^3$ with the descriptions of the *RKHS* and its inner products are found in (Benbourhim and Bouhamidi (2008)) or (Cabrera et al. (2013)). Theory to construct new matrix-valued kernels based on previous matrix kernels or univariate kernels can be found in (Micchelli and Pontil (2005)).

In order to illustrate the Bayesian model described in proposition 25 with simulated data, we decided for a practical choice of $\mathcal{K}$:

$$\mathcal{K}(\mathbf{x},\mathbf{z}) = k(\mathbf{x},\mathbf{z})\left(\gamma \mathbf{1}_{d_2} + (1-\gamma) I_{d_2}\right) \text{ with } 0 \leq \gamma < 1, \tag{5.28}$$

where $\mathbf{1}_{d_2} \in \mathcal{M}_{d_2 \times d_2}(\mathbb{R})$ is the matrix formed with 1's and $k(\mathbf{x},\mathbf{z}) = \exp\left(\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2}\right)$ is the scalar Gaussian kernel. Justification for $\mathcal{K}$ having form (5.28) comes from (Caponnetto et al., 2008, Theorem 12); the corresponding *RKHS* $\mathcal{H}_\mathcal{K}$ is *universal*. Any continuous function $\eta : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ can be uniformly approximated by functions in a *universal RKHS* $\mathcal{H}_\mathcal{K}$. Therefore, if we agree that the regression function η is continuous, η can be uniformly approximated by the solution of (5.1). In this sense, with the choice of (5.28), we only need to assume that the regression function $\eta : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ of problem (5.8) is continuous.

Remark 28

Observe that when $\gamma = 0$, $\mathcal{K}$ becomes diagonal and thus, the possible similarity between components of η is not exploited. If $\gamma = 1$ is set, it is equivalent to assume that all components of η are identical and they are explained using the same functions. Σ describes the correlation of the response components and γ is a parameter that control output dependencies included in the model. Each output of the function $\eta = (\eta_1, \cdots, \eta_{d_2})$ is estimated independently if $\mathcal{K}$ is diagonal and the matrix Σ is assumed diagonal.

Remark 29

Once the kernel $\mathcal{K}$ is fixed and the solution η_λ to the minimization problem (5.1) is obtained, the interpretation is provided by Theorem 51. Up to proving that the hypothesis of Theorem 51 are satisfied, η_λ minimizes the functional $\sum_{i=1}^{n}(\mathbf{y}_i-\eta(\mathbf{x}_i))^\intercal\Sigma^{-1}(\mathbf{y}_i-\eta(\mathbf{x}_i))$ (as function of η) subject to the constrain that $\|\eta\|^2_{\mathcal{H}_\mathcal{K}}<\rho(\lambda)$. Thus, η_λ is the function that best interpolate the data in the sense of $\sum_{i=1}^{n}(\mathbf{y}_i-\eta(\mathbf{x}_i))^\intercal\Sigma^{-1}(\mathbf{y}_i-\eta(\mathbf{x}_i))$ and its smoothness is measured by $\|\cdot\|^2_{\mathcal{H}_\mathcal{K}}$. The trade off between smoothness and interpolation is controlled by λ.

Given $\mathcal{K}$ described in (5.28), lets define $\mathbf{M}:\mathbb{R}^+\times\mathcal{M}^{++}_{d_2\times d_2}(\mathbb{R})\to\mathcal{M}_{nd_2\times nd_2}(\mathbb{R})$ as

$$\begin{aligned}\mathbf{M}(\lambda,H)&:=\mathbf{K}_{xz}\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}+\frac{n\lambda}{2}\Psi^{-1}_{\Sigma,k}\\&=\mathbf{K}_{xz}\left[\Psi_{\Sigma,k}\mathbf{K}_{zz}+\mathbf{K}_{zz}\Psi_{\Sigma,k}\right]^{+}\mathbf{K}_{zx}+\frac{n\lambda}{2}\left(I_n\otimes\Sigma^{-1}\right),\end{aligned}$$

H is hidden in $\mathbf{K}_{xz}$, $\mathbf{K}_{zz}$ and $\mathbf{K}_{zz}$. With the usual constrain $H\in\mathcal{M}^{++}_{d_2\times d_2}(\mathbb{R})$ positive definite and $\lambda=1$, or $\lambda>0$ and $H=I_{d_2}$. A practical problem is the inversion of $\mathbf{M}$. $\mathbf{M}^{-1}$ is needed for the valuation of the scores U, V or M and for the computation of some of the full conditional distributions. It is numerically unstable to directly invert $\mathbf{M}(\lambda,H)$, instead, Proposition 80 provides a more stable expression to compute:

$$\mathbf{M}(\lambda,H)^{-1}=\frac{2}{n\lambda}\Psi_{\Sigma,n}\left[\Psi_{\Sigma^{-1},n}-\mathbf{K}_{xz}\left(\frac{n\lambda}{2}\mathbf{K}_{zz}+\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\mathbf{K}_{zx}\right]\Psi_{\Sigma,n}. \tag{5.29}$$

Observe that the inversion problem of $\mathbf{M}(\lambda,H)$ has been reduced from inverting a $nd_2\times nd_2$ matrix to computing the Moore-Penrose pseudo inverse of a $kd_2\times kd_2$ matrix, with $k<<n$. Even when there could be more than one pseudo inverse of $\frac{n\lambda}{2}\mathbf{K}_{zz}+\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}$. Expression (5.29) is well defined as shown in Proposition 80.

As example of the proposed algorithm's performance, we simulated a training set $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^{100}\subset\mathbb{R}^2\times\mathbb{R}^2$ using model (5.8) with $\eta=(\eta_1,\eta_2)$, η_1 described by (3.17) and η_2 described by (5.25), $\Sigma=\left(\begin{smallmatrix}1 & 0.8\\ 0.8 & 1\end{smallmatrix}\right)$. The sampling points $\{\mathbf{x}_i\}_{i=1}^{100}$ were simulated using $N_2\left(\mathbf{0},\left(\begin{smallmatrix}1 & 0.8\\ 0.8 & 1\end{smallmatrix}\right)\right)$.

The posterior distribution of the parameters from model in proposition 25 and the posterior predictive distribution $\Pi(\eta(\chi)|\mathbf{Y})$ have not an analytically form. Realizations of the posterior distributions can be obtained using Gibss Sampling (Gelman et al., 2014, p. 276-278) by simulating sequentially from the full conditional posterior distributions. Draws $\left\{[\eta(\chi)]_{(i)}\right\}_{i=1}^{N}$

from the posterior predictive $[\eta(\chi)|\mathbf{Y}]$ can be achieved using N draws $\left\{[\mathcal{A}]_{(i)}\right\}_{i=1}^{N}$ from the marginal posterior $[\mathcal{A}|\mathbf{Y}]$ as $[\eta(\chi)]_{(i)} = \sum_{j=1}^{k} \mathcal{K}(\mathbf{z}_j, \chi)[\mathbf{a}_j]_{(i)}$, $[\mathcal{A}]_{(i)} = \text{vec}\left([\mathbf{a}_1]_{(i)} \cdots [\mathbf{a}_k]_{(i)}\right)$ and $[\mathbf{a}_j]_{(i)} \in \mathcal{M}_{d_2 \times 1}(\mathbb{R})$. Three independent chains with different initial over dispersed values, generated randomly, for each parameter were drawn using the MCMC method. Each of the chains were run for $20,000$ iterations discarding the first $15,000$ realizations as warm up and thinning the rest of the sequences by keeping every 5 draws. Convergence Geweke test (Geweke et al. (1991)) and Gelman test (Gelman et al., 2014, p. 285) were used to check the convergence of the chains, for each parameter separately. We use $\hat{\eta}(\chi) := \mathbb{E}\left[\Pi(\eta(\chi)|\mathbf{Y})\right]$ the point estimates of $\eta(\chi)$ and $\tilde{\eta}(\chi) := \text{sd}\left[\Pi(\eta(\chi)|\mathbf{Y})\right]$ the standard deviation of the posterior distribution. $\hat{\eta}(\chi)$ and $\tilde{\eta}(\chi)$ are estimated with the sample mean and the unbiased sample standard deviation from the realizations of the posterior predictive distributions.

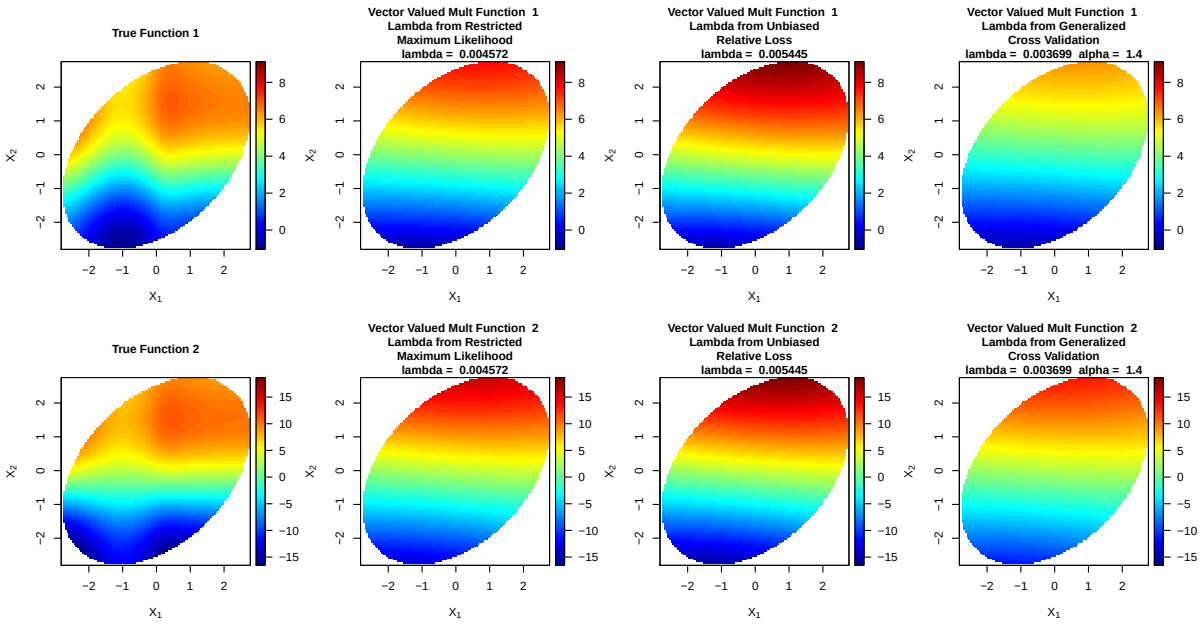

Figure 5.3 Level curves for the true function $\eta = (\eta_1, \eta_2)$ with rank in $\mathbb{R}^2$ (left column). Point Bayes estimates for $\hat{\eta}_1(\chi)$ in first row, second row correspond to estimates of $\hat{\eta}_2(\chi)$. Estimation using the Bayes model from proposition 30, kernel $\mathcal{K}$ described by (5.28) and using a univariate smoothing parameter $\lambda > 0$. Plots in columns 2, 3 and 4 corresponds to estimation using different method to select the bandwidth parameters. Simulated $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{100}$ using model (5.8) with $\eta = (\eta_1, \eta_2)$, η_1 described by (3.17) and η_2 described by (5.25), $\Sigma = \left(\begin{smallmatrix} 1 & -0.8 \\ -0.8 & 1 \end{smallmatrix}\right)$. The sampling points $\{\mathbf{x}_i\}_{i=1}^{100}$ were simulated using $N_2\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$.

Figure 5.3 shows contour levels of an example of prediction from the simulated data using a fitted model described in Proposition 25 and implementation details just explained. The prediction of η is done in a grid $\{\chi_i\}_{i=1}^N \subset \mathbb{R}^2$ of resolution 0.05×0.05 inside and ellipse that would contain 85% of the sampling points simulated from $N_2\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$.

5.3 Bayesian Model for Regression with Measurement Error in the Covariates

Finally, we can consider the vector valued regression problem with classical errors in the covariates. Building over the results developed in this chapter, the next proposition is stated.

Proposition 30

Let $\mathcal{H}$ be a RKHS of functions with domain $\mathbb{R}^{d_1}$ and rank in $\mathbb{R}^{d_2}$. Let $\mathcal{K}$ the reproducing kernel of $\mathcal{H}$. Consider the labeled pairs $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$. We do not observe $\{\mathbf{x}_i\}_{i=1}^n$, instead, we have available noisy repeated measures $\{\{\mathbf{w}_{ij}\}_{j=1}^{n_w}\}_{i=1}^n$ (with out lost of generality n_w does not depend on i). Let $\mathbf{Z} := \{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n =: \mathbf{X}$ be a set of knots with same distribution as the sampling points $\mathbf{X}$, let $\mathbf{Y} = \operatorname{vec}(\mathbf{y}_1\, \mathbf{y}_2 \cdots \mathbf{y}_n)$ the responses, $\mathcal{A} = \operatorname{vec}(\mathbf{a}_1\, \mathbf{a}_2 \cdots \mathbf{a}_k)$, $\mathbf{a}_i \in \mathbb{R}^{d_2}$, let $\mathbf{K}_{xz} \in \mathcal{M}_{nd_2 \times kd_2}(\mathbb{R})$ be a block matrix with (i,j)th block $\mathcal{K}(\mathbf{x}_i, \mathbf{z}_j)$, let $\mathbf{K}_{zz} \in \mathcal{M}_{kd_2 \times xd_2}(\mathbb{R})$ be another block matrix with i,jth block $\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)$, $\mathbf{K}_{zx} = \mathbf{K}_{xz}^\intercal$, $\Psi_{\Sigma,n} = \left(I_n \otimes \Sigma^{-1}\right)$, $\mathbf{\Gamma} = \Psi_{\Sigma,k}\mathbf{K}_{zz} + \mathbf{K}_{zz}\Psi_{\Sigma,k}$. Let $\Sigma \in \mathcal{M}_{d_2 \times d_2}^{++}(\mathbb{R})$ be known or at least we have available a frequentist unbiased estimator $\hat{\Sigma}$ with small variability. Consider the model

$$
\begin{aligned}
\mathbf{y}_i &= \eta(\mathbf{x}_i) + \epsilon_i, \\
\eta(\mathbf{x}) &= \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i, \mathbf{x})\mathbf{a}_i, \\
\mathbf{w}_{ij} &= \mathbf{x}_i + \delta_{ij}, \\
\epsilon_i &\overset{iid}{\sim} N_{d_2}(\mathbf{0}, \Sigma), \\
\delta_{ij} &\overset{iid}{\sim} N_{d_2}(\mathbf{0}, \Sigma_w),
\end{aligned}
$$

with the priors on the parameters

$$\mathcal{A}|\Sigma, H, \lambda, \mathbf{X} \sim N_{kd_2}\left[\mathbf{0}, \frac{2}{n\lambda}\mathbf{\Gamma}^{+}\right]$$

$$\mathbf{P}(\lambda \geq \lambda_0|\Sigma = S, \mathbf{X} = \mathbf{x}) = \int_{\mathbb{R}^{nd_2}} \mathbf{1}\left\{\lambda_0 \geq \underset{x>0, H_1 = I_{d_2}}{\arg\min}\ U(x, H_1|S, \mathbf{y}, \mathbf{x})\right\} dF_{\mathbf{y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}), \quad (5.30)$$

$$\mathbf{x}_i|\boldsymbol{\mu}_x, \Sigma_x \overset{iid}{\sim} N_d(\boldsymbol{\mu}_x, \Sigma_x),\ i = 1, .., n,$$

$$\boldsymbol{\mu}_x|\Sigma_x \sim N_d(\mathbf{d}_x, m_x^{-1}\Sigma_x),$$

$$\Sigma_x \sim Inv - Wishart(\mathbf{A}_x, b_x),$$

$$\Sigma_w \sim Inv - Wishart(\mathbf{A}_w, b_w),$$

$$\mathcal{A}|\mathbf{X} \perp \mathbf{c}|\mathbf{X}, \quad \mathcal{A}|\mathbf{X} \perp (\epsilon_1 \cdots \epsilon_n)^{\intercal}|\mathbf{X}.$$

where $\mathbf{\Gamma}^{+}$ *is the Moore-Penrose inverse of the matrix* $\mathbf{\Gamma}$*. Unless stated otherwise independence is assumed. Alternatively, instead of assigning the prior* (5.30) *to* λ *that depend on* U*, one can assign similar priors but using the function scores* V *of* M*.*

Then the joint posterior of the parameters exists and the full conditional posteriors are

- $[\mathcal{A}|\mathbf{Y}, \mathbf{X}, \mathbf{W}, \lambda, (H = I_{d_2})] \sim N_{kd_2}[\mu_{\mathbf{Y}}, \mathbf{\Sigma}_{\mathbf{Y}}]$*, where*

$$\mu_{\mathbf{Y}} = \mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}\mathbf{Y},$$

$$\mathbf{\Sigma}_{\mathbf{Y}} = \frac{n\lambda}{2}\left\{\mathbf{\Gamma}^{+} - \mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\right\}$$

- $\lambda|\mathbf{X}, \mathbf{Y}, \mathbf{W}, \Sigma = \arg\min_{x>0, H=I_{d_2}}\{V(x, H|\Sigma, \mathbf{Y})\}$ *almost surely.*

  If the other priors on λ *depending on* U *or* M *were assigned, the corresponding full conditional posterior distributions are:*

$$\lambda|\mathbf{X}, \mathbf{Y}, \mathbf{W}, \Sigma = \underset{x>0, H=I_{d_2}}{\arg\min}\ \{U(x, H|\Sigma, \mathbf{Y})\}\ \textit{almost surely},$$

$$\lambda|\mathbf{X}, \mathbf{Y}, \mathbf{W}, \Sigma = \underset{x>0, H=I_{d_2}}{\arg\min}\ \{M(x, H|\Sigma, \mathbf{Y})\}\ \textit{almost surely},$$

- *Conditionally on every parameter (except by* $\mathbf{X}$*),* $\mathbf{x}_i$ *and* $\mathbf{x}_j$ *are independent. And their full conditional distribution is proportional to the next expression,*

$$\begin{aligned}\mathbf{x}_i|\mathbf{Y},\mathbf{W},\mathcal{A},\Sigma_w,\Sigma_x &\propto \left[\mathbf{y}_i|,\sigma^2,\mathbf{x}_i\right]\prod_{j=1}^{n_w}\left[\mathbf{w}_{ij}|\mathbf{x}_i,\Sigma_w\right]\left[\mathbf{x}_i|\boldsymbol{\mu}_x,\Sigma_x\right]\\ &=\exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{y}_i-\eta_{\mathcal{A}}(\mathbf{x}_i)\|^2-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_x)'\Sigma_x^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_x)\right\}\\ &\quad\times\exp\left\{-\frac{1}{2}\sum_{j=1}^{n_x}(\mathbf{w}_{ij}-\mathbf{x}_i)'\Sigma_w^{-1}(\mathbf{w}_{ij}-\mathbf{x}_i)\right\},\end{aligned}$$

- $\Sigma_w|\mathbf{Y},\mathbf{X},\mathbf{W},\Sigma_x \sim Inv-Wishart\left[\mathbf{A}_w+\boldsymbol{A}\boldsymbol{A}',nn_x+b_w\right]$
  where $\mathbf{A}=[\mathbf{x}_1-\mathbf{w}_{11}..\mathbf{x}_1-\mathbf{w}_{1n_x}\cdots\mathbf{x}_n-\mathbf{w}_{n1}..\mathbf{x}_n-\mathbf{w}_{nn_x}]$,

- $\boldsymbol{\mu}_x|\mathbf{Y},\Sigma_x,\mathbf{X}\sim N_d\left[\Sigma_x^{-1}\left(n\bar{\mathbf{x}}+m_x\mathbf{d}_x\right),\frac{1}{n+m_x}\Sigma_x\right]$, *with* $\bar{\mathbf{x}}=n^{-1}\sum_{i=1}^n\mathbf{x}_i$,

- $\Sigma_x|\mathbf{Y},\mathbf{X}=Inv-Wishart\left[\mathbf{A}_x+n\mathbf{S}+\frac{nm_x}{n+m_x}(\bar{\mathbf{x}}-\mathbf{d}_x)(\bar{\mathbf{x}}-\mathbf{d}_x)',n+b_x\right]$
  where $\mathbf{S}=n^{-1}\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})'$.

Furthermore, if $\mu_{\mathbf{Y}}$ *is arranged as* $\mu_{\mathbf{Y}}=\text{vec}\left(\hat{\mathbf{a}}_1\cdots\hat{\mathbf{a}}_k\right)$ *then*

$$\hat{\eta}(\mathbf{x})=\sum_{i=1}^k\mathcal{K}(\mathbf{z}_i,\mathbf{x})\hat{\mathbf{a}}_i \tag{5.31}$$

solves the minimization problem (5.1) *in* $\mathcal{H}_{\mathcal{K}}^{\star}$*, and as* $k\to n$*,* (5.31) *solves* (5.1) *in* $\mathcal{H}_{\mathcal{K}}$.

Proof.

The joint posterior distribution exists because all priors are proper. The full conditional posteriors are obtained by explicitly writing the posterior distribution,

$$\begin{aligned}[\mathcal{A},\,\lambda,\,\Sigma_x,\,\Sigma_w,\,\mathbf{X},\mu_x|\mathbf{Y},\,\mathbf{W},\Sigma] &= [\mathbf{Y}|\mathcal{A},\,\mathbf{X},\,\lambda,\,\Sigma]\times[\mathbf{W}|\mathbf{X},\Sigma_w]\\ &\quad\times[\mathcal{A}|\mathbf{X},\,\Sigma,\,\lambda]\times[\mathbf{X}|\mu_x,\Sigma_x]\times[\mu_x|\Sigma_x]\times[\Sigma_x]\times[\Sigma_w],\end{aligned}$$

and noticing the factors that correspond to each distribution and completing terms. The only difficult derivation is $[\mathcal{A}|\mathbf{X},\,\Sigma,\,\lambda]$ and it was obtained in Proposition 25.

In Section 5.1 was argued that a sufficient condition for a vector of the form $\mu_{\mathbf{Y}}=\text{vec}\left(\hat{\mathbf{a}}_1\cdots\hat{\mathbf{a}}_k\right)$ to induce a solution to (5.1) in $\mathcal{H}_{\mathcal{K}}$ is that $\mu_{\mathbf{Y}}$ satisfies equation (5.7). If such condition is satisfied then (5.31) would be the unique solution to (5.1). In Proposition 79 we

state and prove that $\mu_{\mathbf{Y}}$ satisfies such condition, therefore (5.31) is the unique solution to the functional minimization problem. ■

Remark 31

Conditionally on $\mathbf{X}$*, model in Proposition 30 has the same form as the model for the measurement free error case in Proposition 25 and same full conditional posterior in the common parameters. As consequence of these remarks, interpretation of the conditional fitted regressions are the same for both models in such propositions; the mean of the full conditional posteriors predictive* $\hat{\eta}(\chi) = [\eta(\chi)|bX, \mathbf{Y}, \lambda, H, \mathbf{\Sigma}]$ *as a deterministic function of* $\chi \in \mathbb{R}^{d_1}$ *is the solution to (5.1) in span*$\{\mathcal{K}(\mathbf{z}_i, \cdot), i = 1, ..., k\}$*. As* $k \to n$*,* $\hat{\eta}$ *solves (5.1) in* $\mathcal{H}_{\mathcal{K}}$*.*

In the multivariate real regression problem with measurement errors in the covariates, the variance error σ^2 of the response is dramatically overestimated when using the proposed models from previous chapters. The overestimations of σ^2 do not directly affect the choices of the smoothing parameters when using the GCV method or the MRL method and adequate estimates of η are achieved, this is because such scores do not depend on the variance. The overestimations of σ^2 in the Bayesian procedure, lowers the empirical coverages of the credible intervals when using the UERL method to select the smoothing parameters; this is because the score UERL $\mathcal{U}$ depend on σ^2, although the point estimates of η are still acceptable. The same problem arrives in the current setting of vector-valued regression. Estimating Σ, either with a difference-based method inside the full model of Proposition 30 (as discussed in Section 4.2 and assigning priors in a similar way as prior (4.12); or either by assigning a, prior such as inverse-Wishart in Proposition 25, results in estimators that provide large variances $\hat{\Sigma}_{ii}$ and misleading pairwise covariances $\hat{\Sigma}_{ij}$. If η and Σ are attempted to be jointly estimated using a Bayesian model such as Proposition 30, the three proposed methods for selection of bandwidths provide λ too large as an effect of the dependence of Σ, resulting in an over-smoothed function. This effect was observed in all our simulated examples. Thus the reason we request in Proposition 30 for a known Σ, or a frequentist unbiased estimator with small variability. It is an open problem to estimate η and Σ at the same time in presence of measurement error.

As in the univariate case, if $n_w = 50$, the problem we just described is minimized and we are still able to adequately estimate η and Σ by assigning an inverse Wishart prior to Σ.

In regard to the prior proposed for the latent variable $\mathbf{x}_i$'s, we agree that there is not a reasonable distribution in the general case. The prior may be changed according to the required application; for example, a flat reference prior may be reasonable; normal hierarchical could be another option, mixture of normal distributions is a flexible prior but in this case we may have problems identifying the group of the mixture to where and observation belong Gelman et al. (2014), and the updates in the MCMC algorithm may be slow. We used a hierarchical normal prior.

5.3.1 Implementation and example of estimation

A training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{100} \subset \mathbb{R}^2 \times \mathbb{R}^2$ was simulated using the model in proposition 30 with $\eta = (\eta_1, \eta_2)$, η_1 described by (3.17) and η_2 described by (5.25), $\Sigma = \left(\begin{smallmatrix} 1 & -0.8 \\ -0.8 & 1 \end{smallmatrix}\right)$. The sampling points $\{\mathbf{x}_i\}_{i=1}^{100}$ were simulated using $N_2\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$ and the errors $\{\{\delta_{i,j}\}_{j=1}^{n_w}\}_{i=1}^{100} = N_2\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$, $n_w = 7$. Figure 5.3 shows contour levels of an example of prediction from a fitted model using simulated data. The model to consider is described in Proposition 25, and the required matrix computation is straightforward with the exception of the inversion of the matrix $\mathbf{M}(\lambda, H)$; expression (5.29) provides an explicit computationally more stable inversion of $\mathbb{M}(\lambda, H)$. The prediction of η is done in a grid $\{\chi_i\}_{i=1}^{N} \subset \mathbb{R}^2$ of resolution 0.05×0.05 inside an ellipse that would contain 85% of the sampling points simulated from $N_2\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$.

The reproducing kernel (5.28) is used as basis function for the implementation of the algorithm. The size of the sets of knots $\{\mathbf{z}_i\}_{i=1}^{k}$ is chosen to be

$$k = \lfloor \max\left\{30, 10n^{2/9}\right\} \rfloor$$

as described in Section 2.2.3. We use the averages $\left\{n_w^{-1}\sum_{j=1}^{n_w} \mathbf{w}ij\right\}_{i=1}^{n}$ as if they were the true latent variables only when the knots are to be computed. The algorithm to select the knots is described in Section 2.2.3.

As fitting procedure, we use a Metropolis-Hasting within Gibbs sampler. Simulations from the full conditional posterior of $\mathbf{x}_i$'s are achieved independently using a Metropolis-Hasting

step. We use the *adaptive Metropolis* algorithm proposed by Roberts and Rosenthal (2009). At the i^{th} iteration the proposal distribution for $i \leq 2d$ is

$$Q_i(\mathbf{x}) = N(\mathbf{x}, 0.1^2 \mathbf{1}_d c_i/d);$$

for $i > 2d$ we use

$$Q_i(\mathbf{x}) = \begin{cases} (1-\theta)N(\mathbf{x}, 2.38^2 \tilde{\Sigma}_i c_i/d) + \theta N(\mathbf{x}, 0.1^2 \mathbf{1}_d c_i/d) & \tilde{\Sigma}_i > 0, \\ N(\mathbf{x}, 0.1^2 \mathbf{1}_d c_i/d) & \tilde{\Sigma}_i \not> 0, \end{cases}$$

where $\theta \in (0,1)$ and the empirical covariance matrix $\tilde{\Sigma}_i$ is described with

$$\tilde{\Sigma}_i = \frac{1}{i}\left(\sum_{j=0}^{i} \mathbf{x}_j \mathbf{x}_j^\intercal - (i+1)\bar{\mathbf{x}}_i \bar{\mathbf{x}}_i^\intercal\right)$$

where $\bar{\mathbf{x}}_i = \frac{1}{i+1}\sum_{j=0}^{i} \mathbf{x}_j$, $c_i = exp\left(min(1^{-2}, n^{-\frac{1}{2}})(\mathbf{1}_{accpt_i > .44} - \mathbf{1}_{accpt_i < .44})\right)$. We use $\theta = .5$.

Three independent chains with different initial over dispersed values for each parameter were drawn. $20,000$ iterations are computed for each chain and the first $15,000$ are discarded as burn-in. The chains are thinned keeping every 5 iterations. For testing convergence of the MCMC chains of the parameters we use Gelman-Rubin test (Gelman et al., 2014, p. 285) and Geweke test, (Geweke et al. (1991)) for the chains of each parameter. The tests were applied independently for each parameter.

For $\chi \in \mathbb{R}^d$ let $\hat{\eta}(\chi) := \mathbb{E}\left[\Pi(\eta(\chi)|\mathbf{Y}, \mathbf{W})\right]$ the point estimates of $\eta(\chi)$. $\hat{\eta}(\chi)$ is estimated using the sample mean from the realizations of the posterior predictive distributions. An example of the fitted regression function from simulated data is displayed in Figure 5.4. The functions V, U and M depending only on $\lambda > 0$ and $H = I_{d_2}$ has been observed to have a global minimum most of the time in the interval $(10^{-4.5}, 10^{-2.5})$, we set this range as bound to search for the minimum in every iteration of the MCMC.

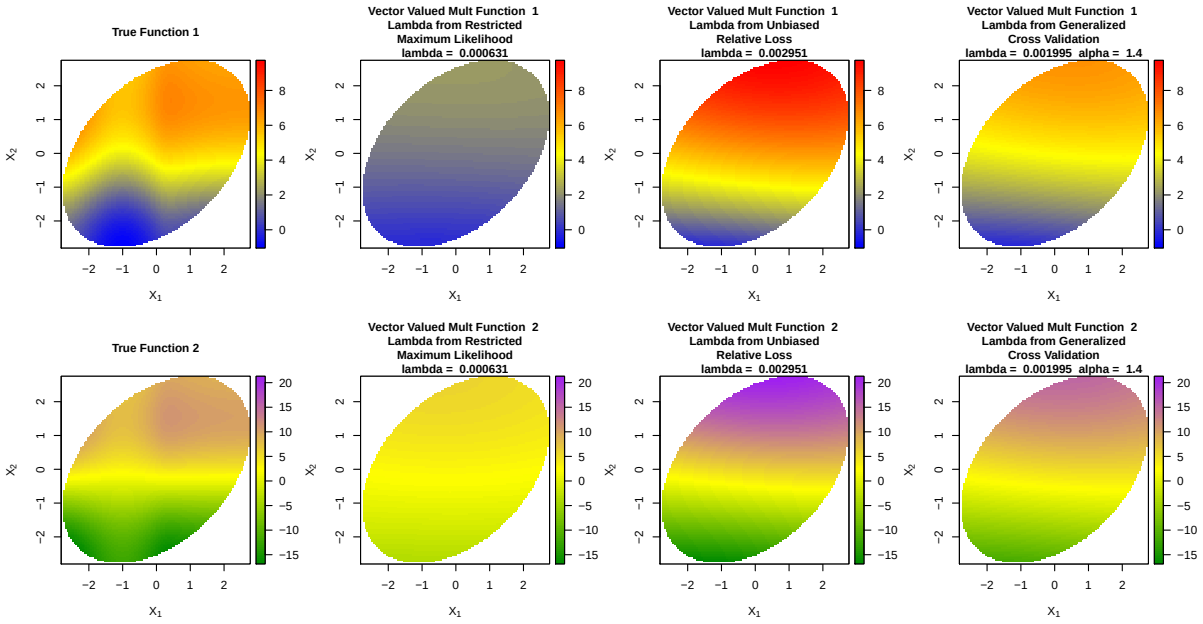

Figure 5.4 Level curves for the true function $\eta = (\eta_1, \eta_2)$ with rank in $\mathbb{R}^2$ (left column). Point Bayes estimates for $\hat{\eta}_1(\chi)$ in first row, second row correspond to estimates of $\hat{\eta}_2(\chi)$. Estimation using the Bayes model from Proposition 30, kernel $\mathcal{K}$ described by (5.28). We used $\lambda > 0$ smoothing parameter. Plots in columns 2, 3 and 4 corresponds to estimation using different method to select the bandwidth parameters. Simulated $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{100}$ was obtained using model (5.8) with $\eta = (\eta_1, \eta_2)$, η_1 described by (3.17) and η_2 described by (5.25), $\Sigma = \left(\begin{smallmatrix} 1 & -0.8 \\ -0.8 & 1 \end{smallmatrix}\right)$. The sampling points $\{\mathbf{x}_i\}_{i=1}^{100}$ were simulated using $N_2\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$ and the errors $\{\{\delta_{i,j}\}_{j=1}^{7}\}_{i=1}^{100}$ where simulated from $\delta_{i,j} \overset{iid}{\sim} N_2\left(\mathbf{0}, \left(\begin{smallmatrix} 1 & 0.8 \\ 0.8 & 1 \end{smallmatrix}\right)\right)$.

5.4 Conclusions

Bayesian regression models were proposed for a vector valued multivariate regression problem using labeled data. The models incorporate theoretical frequentist results from learning a multi-output function using kernel methods in a framework of vector-valued reproducing kernel Hilbert spaces. The innovation of the proposed models is that the Bayesian machinery allows for credible regions of estimation, and point Bayes estimates of the regression function conserve similar properties as the frequentist multi-output kernel regressions. Furthermore, we proposed three methods to select the real bandwidth parameter. Such methods are conjectured to be extensions of the Unbiased Estimate Relative Loss method, Generalized Cross Validation score

and the RML method using Bayes model described in previous chapters. Theoretical extensions of the real smoothing parameter methods are proposed for the case of bandwidth matrices.

Furthermore, we propose a Bayesian model to estimate vector valued multivariate regression functions using labeled data with classical measurement errors in the covariates. By observations of the fitted models over simulated data, the models seems to recover the regression function adequately but only when Σ is known or at least a low variability unbiased frequentist estimator is available. Simulated data were generated and the proposed models under these considerations were fitted. Examples are shown and at least for these cases, the regression function seems to be recovered.

BIBLIOGRAPHY

Agarwal, A., Gerber, S., and Daume, H. (2010). Learning multiple tasks using manifold regularization. In *Advances in neural information processing systems*, pages 46–54.

Akhiezer, N. I. and Glazman, I. M. (1981a). *Theory of linear operators in Hilbert space Volume 1.* Boston: Pitman Pub.

Akhiezer, N. I. and Glazman, I. M. (1981b). *Theory of linear operators in Hilbert space Volume 2.* Boston: Pitman Pub.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Athreya, K. B. and Lahiri, S. N. (2006). *Measure theory and probability theory.* Springer Science & Business Media.

Banerjee, S. and Roy, A. (2014). *Linear algebra and matrix analysis for statistics.* CRC Press.

Barry, D. et al. (1986). Nonparametric bayesian regression. *The Annals of Statistics*, 14(3):934–953.

Belkin, M., Niyogi, P., and Sindhwani, V. (2005). On manifold regularization. In *AISTATS.* Citeseer.

Bellman, R. and Dreyfus, S. (1962). Applied dynamic programming.

Benbourhim, M. N. and Bouhamidi, A. (2005). Approximation of vectors fields by thin plate splines with tension. *Journal of Approximation Theory*, 136(2):198–229.

Benbourhim, M.-N. and Bouhamidi, A. (2008). Error estimates for interpolating div–curl splines under tension on a bounded domain. *Journal of Approximation Theory*, 152(1):66–81.

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002). Bayesian smoothing and regression splines for measurement error problems. *Journal of the American Statistical Association*, 97(457):160–169.

Bhatia, R. (2009). *Positive definite matrices*. Princeton University Press.

Bia, M., Van Kerm, P., et al. (2014). Space-filling location selection. *Stata Journal*, 14(3):605–622.

Bilodeau, M. and Brenner, D. (2008). *Theory of multivariate statistics*. Springer Science & Business Media.

Cabrera, D. A. C., Gonzalez-Casanova, P., Gout, C., Juárez, L. H., and Reséndiz, L. R. (2013). Vector field approximation using radial basis functions. *Journal of Computational and Applied Mathematics*, 240:163–173.

Camber, H. A. (1979). Choice of an optimal shape parameter when smoothing noisy data. *Communications in Statistics-Theory and Methods*, 8(14):1425–1435.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Caponnetto, A., Micchelli, C. A., Pontil, M., and Ying, Y. (2008). Universal multi-task kernels. *Journal of Machine Learning Research*, 9(Jul):1615–1646.

Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04):377–408.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association*, 83(404):1184–1186.

Carroll, R. J., Spiegelman, C. H., Lan, K. G., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, 71(1):19–25.

Castro, M., Bolfarine, H., and Galea, M. (2013). Bayesian inference in measurement error models for replicated data. *Environmetrics*, 24(1):22–30.

Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 473–491.

Corwin, L. (1982). *Multivariable calculus*, volume 64. CRC Press.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

De Brabanter, K., Ferrario, P. G., and Györfi, L. (2014). Detecting ineffective features for nonparametric regression. *Regularization, Optimization, Kernels, and Support Vector Machines*, pages 177–194.

Delaigle, A. (2014). Nonparametric kernel methods with errors-in-variables: Constructing estimators, computing them, and avoiding common mistakes. *Australian & New Zealand Journal of Statistics*, 56(2):105–124.

Delaigle, A. and Hall, P. (2011). Estimation of observation-error variance in errors-in-variables regression. *Statistica Sinica*, pages 1023–1063.

Devroye, L., Schäfer, D., Györfi, L., and Walk, H. (2003). The estimation problem of minimum mean squared error. *Statistics & Decisions/International mathematical Journal for stochastic methods and models*, 21(1/2003):15–28.

Douglas Nychka, Reinhard Furrer, John Paige, and Stephan Sain (2015). fields: Tools for spatial data. R package version 8.4-1.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer.

Fan, J. (1990). Asymptotic normality for deconvolving kernel density estimators.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pages 1257–1272.

Fristedt, B. E. and Gray, L. F. (2013). *A modern approach to probability theory*. Springer Science & Business Media.

Fuller, W. A. (2009). *Measurement error models*, volume 305. John Wiley & Sons.

Geary, R. C. (1941). Inherent relations between random variables. In *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences*, pages 63–76. JSTOR.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014). *Bayesian data analysis*, volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media.

Gu, C. (2014). Smoothing spline anova models: R package gss. *Journal of Statistical Software*, 58(5):1–25.

Gu, C. and Kim, Y.-J. (2002). Penalized likelihood regression: general formulation and efficient approximation. *Canadian Journal of Statistics*, 30(4):619–628.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234.

Gu, C. and Wahba, G. (1993a). Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 353–368.

Gu, C. and Wahba, G. (1993b). Smoothing spline anova with component-wise bayesian "confidence intervals". *Journal of Computational and Graphical Statistics*, 2(1):97–117.

Hausman, J. (2001). Mismeasured variables in econometric analysis: problems from the right and problems from the left. *The Journal of Economic Perspectives*, 15(4):57–67.

Henderson, C. R. (1950). Estimation of genetic parameters. In *Biometrics*, volume 6, pages 186–187. INTERNATIONAL BIOMETRIC SOC 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, 1973(Symposium):10–41.

Hoffman, K. and Kunze, R. (1990). Linear algebra, 2nd.

Kim, Y.-J. and Gu, C. (2004). Smoothing spline gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):337–356.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Königsberger, K. (2013). *Analysis 2*. Springer-Verlag.

Kurdila, A. J. and Zabarankin, M. (2006). *Convex functional analysis*. Springer Science & Business Media.

Li, K.-C. (1986). Asymptotic optimality of cl and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, pages 1101–1112.

Liitiäinen, E., Corona, F., and Lendasse, A. (2008). On nonparametric residual variance estimation. *Neural Processing Letters*, 28(3):155–167.

Liitiäinen, E., Corona, F., and Lendasse, A. (2010). Residual variance estimation using a nearest neighbor statistic. *Journal of Multivariate Analysis*, 101(4):811–823.

Liitiäinen, E., Verleysen, M., Corona, F., and Lendasse, A. (2009). Residual variance estimation in machine learning. *Neurocomputing*, 72(16):3692–3703.

Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics*, 17(4):427–438.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 30(2):292–304.

Micchelli, C. A. and Pontil, M. (2004). Kernels for multi–task learning. In *NIPS*, volume 86, page 89.

Micchelli, C. A. and Pontil, M. (2005). On learning vector-valued functions. *Neural computation*, 17(1):177–204.

Minh, H. Q., Bazzani, L., and Murino, V. (2013). A unifying framework for vector-valued manifold regularization and multi-view learning. In *ICML (2)*, pages 100–108.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association*, 83(404):1134–1143.

Penrose, R. (1955). A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge Univ Press.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, 81(394):321–327.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica: Journal of the Econometric Society*, pages 375–389.

Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics*, 18(2):349–367.

Robinson, G. K. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, pages 15–32.

Royle, J. A. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in splus. *Computers & Geosciences*, 24(5):479–488.

Rudin, W. (1991). Functional analysis. international series in pure and applied mathematics.

Ruppert, D. (2012). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, 74(2):385–391.

Schennach, S. M., Hu, Y., Lewbel, A., et al. (2007). Nonparametric identification of the classical errors-in-variables model without side information. *Working Papers in Economics*, page 426.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer.

Spiriti, S., Eubank, R., Smith, P. W., and Young, D. (2013). Knot selection for least-squares and penalized splines. *Journal of Statistical Computation and Simulation*, 83(6):1020–1036.

Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *Journal of Multivariate Analysis*, 82(1):111–133.

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, 72(3):583–592.

Stefanski, L. A. and Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, pages 1335–1351.

Stefanski, L. A. and Carroll, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, 74(4):703–716.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics*, 21(2):169–184.

Stewart, J. (2011). *Multivariable calculus*. Cengage Learning.

Tong, T., Ma, Y., Wang, Y., et al. (2013). Optimal variance estimation without estimating the mean function. *Bernoulli*, 19(5A):1839–1854.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 364–372.

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *The annals of statistics*, pages 146–156.

Wahba, G. (1983). Bayesian" confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 133–150.

Wahba, G. (1985). A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, pages 1378–1402.

Wahba, G. (1987). Partial and interaction spline models for the semiparametric estimation of functions of several variables.

Wahba, G. and Craven, P. (1978). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–404.

Wahba, G. and Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly weather review*, 108(8):1122–1143.

Wecker, W. E. and Ansley, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, 78(381):81–89.

Weidmann, J. (1980). Linear operators in hilbert spaces.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114.

Yang Xiang, Gubian, S., Suomela, B., and Hoeng, J. (2013). Generalized simulated annealing for efficient global optimization: the GenSA package for R. *The R Journal Volume 5/1, June 2013*.

Zhou, Y., Cheng, Y., Wang, L., and Tong, T. (2015). Optimal difference-based variance estimation in heteroscedastic nonparametric regression. *Statistica Sinica*, 25:1377–1397.

APPENDIX A. DEFINITIONS

Definition 32 (Moore-Penrose Inverse of a Matrix)

Let $\mathfrak{K}$ be a field ($\mathbb{R}$ or $\mathbb{C}$) and $M \in \mathcal{M}_{n\times m}(\mathfrak{K})$. The Moore-Penrose Inverse $M^+ \in \mathcal{M}_{m\times n}(\mathfrak{K})$ is the unique matrix, Penrose (1955), satisfying all of the next criteria:

$$AA^+A = A,$$

$$A^+AA^+ = A^+,$$

$$\left(AA^+\right)^* = AA^+,$$

$$\left(A^+A\right)^* = A^+A,$$

where A^* denotes the Hermitian transpose ($A^* = B^\intercal$ if $\mathfrak{K} = \mathbb{R}$). If $A \in \mathcal{M}_{n\times n}(\mathfrak{K})$ is invertible then $A^+ = A^{-1}$.

Definition 33 (Square Root of a Square Matrix)

Let $M \in \mathcal{M}_{n\times n}(\mathbb{R})$ with singular value decomposition (Banerjee and Roy (2014))

$$M = Udiag(\lambda_1 \cdots \lambda_n)V^\intercal$$

where $U, V \in \mathcal{M}_{n\times n}(\mathbb{R})$ are orthogonal matrices and $\lambda_i \geq 0$. The square root matrix of M is

$$M^{\frac{1}{2}} := Udiag(\sqrt{\lambda_1} \cdots \sqrt{\lambda_n})V^\intercal.$$

Definition 34 (Complete Linear Space)

A linear space $\mathcal{H}$ is complete if every Cauchy sequence converges to an element in $\mathcal{H}$.

Definition 35 (Hilbert Space)

A Hilbert space $\mathcal{H}$ is a complete inner product linear space.

Definition 36 (Banach Space)

A Banach space $\mathcal{B}$ is vector space over $\mathbb{R}$ (or over $\mathbb{C}$). Such space has a norm and is complete with respect to that norm.

Definition 37 (Self-adjoint Bounded Linear Operator)

Let $\mathbb{Y}$ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathbb{Y}}$, let $\mathfrak{L}(\mathbb{Y})$ be the Banach space of bounded linear operators on $\mathbb{Y}$, let $A \in \mathfrak{L}(\mathbb{Y})$ a linear operator. By the Riez Representation Theorem (Rudin, 1991, § 12.9) there exists $A^\star \in \mathfrak{L}(\mathbb{Y})$ such that for any $\mathbf{x}, \mathbf{z} \in \mathbb{Y}$, we have $\langle A\mathbf{x}, \mathbf{z} \rangle_{\mathbb{Y}} = \langle \mathbf{x}, A^\star \mathbf{z} \rangle_{\mathbb{Y}}$. The adjoint of A is $A^\star$. A is a self-adjoint linear operator if $A = A^\star$.

Definition 38 (Separable Hilbert Space)

A Hilbert space $\mathcal{H}$ is separable if there exist a dense countable subset. That is to say that $\mathcal{H}$ is separable iff there exists $\{\mathbf{x}_i\}_{i=1}^{\infty} \subset \mathcal{H}$ such that every non empty subset of $\mathcal{H}$ contains at least one element of $\{\mathbf{x}_i\}_{i=1}^{\infty}$.

Definition 39 (Orthogonality)

Let $\mathcal{H}$ be a Hilbert space with inner product $(\cdot, \cdot)$.

- *$\eta, \zeta \in \mathcal{H}$ are orthogonal iff $(\eta, \zeta) = 0$ and is denoted as $\eta \perp \zeta$.*
- *Two subsets $\mathcal{G}_1, \mathcal{G}_2 \subset \mathcal{H}$ are orthogonal (in symbols $\mathcal{G}_1 \perp \mathcal{G}_2$) iff $(\eta_1, \eta_2) = 0, \forall \eta_1 \in \mathcal{G}_1$ and $\forall \eta_2 \in \mathcal{G}_2$.*
- *For $\mathcal{G} \subset \mathcal{H}$, the set $\mathcal{G}^\perp := \{\eta \in \mathcal{H} : \{\eta\} \perp \mathcal{G}\}$ is the orthogonal complement of $\mathcal{G}$.*

Definition 40 (Projection onto Closed Subspace)

Let $\mathcal{H}$ a Hilbert space and $\mathcal{G} \subset \mathcal{H}$ a closed subspace. Let $f \in \mathcal{H}$, by Theorem 47 it can be uniquely decomposed as $f = g + h$ with $g \in \mathcal{G}$ and $h \in \mathcal{G}^\perp$. The projection of f onto $\mathcal{G}$ is g.

Definition 41 (Tensor Sum Decomposition)

Let $\mathcal{H}$ a Hilbert space and $\mathcal{G} \subset \mathcal{H}$ a closed subspace. The unique decomposition of any $f \in \mathcal{H}$ as $f = g + h$ with g the projection of f onto $\mathcal{G}$ and $h = f - g \in \mathcal{G}^\perp$ defines the tensor sum decomposition of $\mathcal{H}$ denoted as $\mathcal{H} = \mathcal{G} \bigoplus \mathcal{G}^\perp$. Multiple term tensor decomposition is defined recursively

Definition 42 (Gâteaux derivative)

Let A be a functional in a linear space $\mathfrak{L}$. For $f, g \in \mathfrak{L}$, let $A_{f,g}(\alpha) := A(f + \alpha g)$ a function of $\alpha \in \mathbb{R}$. If $\dot{A}_{f,g}(0)$ exists and is linear in g and $\forall f \in \mathfrak{L}$, then A is Gâteaux differentiable in $\mathfrak{L}$, and $\dot{A}_{f,g}(0)$ is the Gâteaux derivative of A at f in the direction of g.

Definition 43 (Inverse Gamma Distribution)

A random variable has inverse gamma distribution with parameter $\alpha > 0$ and $\beta > 0$, notation $Inv - Gamma(\alpha, \beta)$ iff its probability density function is

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{-(\alpha+1)} \exp\left(-\frac{1}{\beta x}\right) \mathbf{1}_{\{x>0\}},$$

where Γ denotes the gamma function.

Definition 44 (k-th nearest neighbor, (De Brabanter et al. (2014)))

Let a pair set $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be available and a metric ρ in $\mathbb{R}^d$. The k-th nearest neighbor of y_i among $\{\mathbf{x}_j\}_{j=1, j\neq k}^n$, denoted as $y_{\{n,i,k\}} \in \{y_i\}_{i=1}^n$ or by simplicity $y_{i,(k)}$ where

$$\{n, i, k\} = \underset{1 \leq j \leq n,\, j \neq i,\, j \notin \{\{n,i,1\}, \cdots \{n,i,k-1\}\}}{\arg\min} \rho(\mathbf{x}_i, \mathbf{x}_j)$$

Definition 45 *Let $\{x_n\}_{i=1}^{\infty}$ a set of random variables defined in the same probability space. Let $\{x_n\}_{i=1}^{\infty} \subset \mathbb{R}$. the notation*

$$x_n = o_p(a_n)$$

describes that x_n/a_n converge to zero in probability.

APPENDIX B. MISCELLANEOUS PROPOSITIONS AND THEOREMS

This appendix is dedicated to declare and prove the auxiliary results used in the main body of the dissertation. The appendix contains theorems, propositions, lemmas and corollaries. Some of the results were found in the literature with or without their respective proof, while most of the propositions are our own work. The rule to differentiate between new result and result found in the literature is by the presence of a formal proof; if the results has a complete proof the auxiliary result is ours, if only the idea of the proof or the reference was made, then the result was found in the literature.

B.1 Real Valued Functions in Hilbert Spaces

Proposition 46

In the context of Proposition 8, $\xi(\mathbf{x}) \in Im(Q)$.

Proof.

Let $\mathbb{A} = \{c : c^\intercal Q c = 0\}$. Let $\mathbf{x} \in \mathbb{R}^l$ and $\mathbf{c} \in \mathbb{A}$.

Observe that $0 = \mathbf{c}^\intercal Q \mathbf{c} = J(\xi(\mathbf{x})^\intercal \mathbf{c})$ implies $\xi(\mathbf{x})^\intercal \mathbf{c} = 0$ because J is a norm in $span\left\{R_J(\mathbf{z}_i, \cdot)\right\}_{i=1}^k = \mathcal{H}^\star \ominus \mathcal{N}_J$.

Then $\mathbf{c} \perp \xi(\mathbf{x})\ \forall \mathbf{x} \in \mathbb{R}^l$ and $\forall \mathbf{c} \in \mathbb{A}$. Therefore $\xi(\mathbf{x}) \in \mathbb{A}^\perp$.

On another side, from definitions we observe that $\ker(Q) \subset \mathbb{A}$, then $Im(Q) = \ker(Q)^\perp \supseteq \mathbb{A}^\perp$; thus concluding that $\xi(\mathbf{x}) \in Im(Q)$ ■

Theorem 47 (Projection Theorem)

Let $\mathcal{H}$ be a Hilbert space and $\mathcal{G} \subset \mathcal{H}$ a closed subspace. Then we have ${\mathcal{G}^\perp}^\perp = \mathcal{G}$. Each $f \in \mathcal{H}$ can be uniquely decomposed in the form $f = g + h$ with $g \in \mathcal{G}$ and $h \in \mathcal{G}^\perp$.

Proof. Theorem 3.2 (Weidmann, 1980, pag. 31). ■

Proposition 48

Let $\mathbb{X}$ be a non-empty set and $H = \{f : \mathbb{X} \to \mathbb{R}\}$ be a Hilbert space with a semi-inner product and induced square semi-norm J. Let $\mathcal{N}_J = \{f | J(f) = 0\}$ the null space of J which is of finite dimension. For given pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathbb{X} \times \mathbb{R}$ define the functional

$$\mathcal{L}(f) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 . \tag{B.1}$$

Then,

i $\mathcal{L}$ is continuous, convex and Gâteaux differentiable.

ii If $\{\phi_i\}_{i=1}^l$ is a basis of $\mathcal{N}_J$ and S matrix of size $n \times l$ such $S_{i,j} = \phi_j(\mathbf{x}_i)$ and full column rank, then $\mathcal{L}$ is strictly convex in $\mathcal{N}_J$.

iii If S is full column rank and $\lambda > 0$, then $\mathcal{L} + \lambda J$ is strictly convex in H.

Proof. Note for me: I have to show in part ii) that the choosing of the basis is not important.

■

Proposition 49 *If $\mathcal{A}$ is a strictly convex functional in a Hilbert space H with a local minimum, then $\mathcal{A}$ has a global minimum.*

Proof.

Let η be a minimum of $\mathcal{A}$ and pick $f \in \mathcal{A}$ where $\eta \neq f$.

By definition of local minimum there must be an open set $U \subset \mathcal{H}$ around η such that $\mathcal{A}(\eta) \leq \mathcal{A}(g), \forall g \in U$. We can take $g = \eta + t(f - \eta) = tf + (1 - t)\eta$ since for small enough $t > 0$ we have $g \in U$ (we use the completeness of $\mathcal{H}$ as well). Then

$$\begin{aligned} \mathcal{A}(\eta) &\leq \mathcal{A}(g) \\ &= \mathcal{A}\left(tf + (1-t)\eta\right) \\ &< t\mathcal{A}(f) + (1-t)\mathcal{A}(\eta) \end{aligned}$$

for small $t > 0$, where the last inequality is because $\mathcal{A}$ is strictly convex.

Then $t\mathcal{A}(\eta) < t\mathcal{A}(f)$ and $\mathcal{A}(\eta) < \mathcal{A}(f)$ follows.

Since $f \in \mathcal{H}$ was arbitrary we have shown that η is a global minimum of $\mathcal{A}$, which is unique.

■

Theorem 50 (Existence of Minimizer) *(Gu and Qiu (1993))*

Suppose L is a continuous and convex functional in a Hilbert space $\mathcal{H}$ and J is a square (semi) norm in $\mathcal{H}$ with a finite dimensional null space $\mathcal{N}_J$. If L has a unique minimizer in $\mathcal{N}_J$, then $L + \lambda J$ has a minimizer in $\mathcal{H}$.

Theorem 51 (Optimization Interpretation) *(Gu, 2013, Theorem 2.12, pag 53)*

Let $\mathcal{H}$ be a Hilbert space with a (semi) inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and $J(f) := \langle f, f \rangle_{\mathcal{H}}$ denote the induced square (semi) norm in $\mathcal{H}$. Let $\mathcal{N}_J = \{f \in \mathcal{H}, J(f) = 0\}$ be the null space of J. Suppose $L : \mathcal{H} \to \mathbb{R}$ is a continuous, convex functional and Gâteaux differentiable. Let $\rho > 0$ and define the functional $\Lambda : \mathcal{H} \to \mathbb{R}$ as $\Lambda(f) := -\rho^{-1} \dot{L}_{f,f_1}(0)$ with f_1 being the projection of f into $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$. Then

(1) If $f^* = \underset{f \in \mathrm{C}_\rho}{\operatorname{argmin}} L(f)$, where $\mathrm{C}_\rho = \{f \in H, J(f) \leq \rho\}$, then

$$f^* = \underset{f \in H}{\operatorname{argmin}} L(f) + \frac{\Lambda(f^*)}{2} J(f).$$

(2) For $\lambda > 0$, $f^0 = \underset{f \in H}{\operatorname{argmin}} L(f) + \frac{\lambda}{2} J(f)$ and $\mathrm{E}_\lambda = \left\{f \in H, J(f) \leq J(f^0)\right\}$, then

$$f^0 = \underset{f \in \mathrm{E}_\lambda}{\operatorname{argmin}} L(f).$$

Theorem 52 (Representer Theorem) *Schölkopf et al. (2001)*

Let $\mathbb{X}$ be a non-empty set, $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{X} \times \mathbb{R}$ be a training set, $R : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a kernel with respective RKHS $\mathcal{H}$ and norm $\| \cdot \|$, $g : [0, \infty) \to \mathbb{R}$ be a non-decreasing function, $\mathrm{C} : (\mathbb{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \infty$ be a cost function, and define set of real valued functions $\{\phi_i\}_{i=1}^l$ on $\mathbb{X}$ with the property that the $n \times l$ matrix $(\phi_i(\mathbf{x}_j))_{i,j}$ is full rank. Then any $f \in span\{\phi_i\} \bigoplus \mathcal{H}$ solving

$$\underset{f \in span\{\phi_i\} \bigoplus \mathcal{H}}{\arg\min} \mathrm{C}\left((\mathbf{x}_1, y_1, f(\mathbf{x}_1)), \cdots, (\mathbf{x}_n, y_n, f(\mathbf{x}_n))\right) + g\left(\|f\|\right)$$

admits a representation of the form

$$f(\cdot) = \sum_{i=1}^{l} d_i \phi_i(\cdot) + \sum_{i=1}^{n} c_i R(\mathbf{x}_i, \cdot)$$

with unique coefficients $\{c_i\}_{i=1}^n \subset \mathbb{R}$*, but not necessarily unique* $\{d_i\}_{i=1}^l \subset \mathbb{R}$*.*

Proposition 53

Let $\mathbb{X}$ *be a non-empty set,* $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1 \subset \{f : \mathbb{X} \to \mathbb{R}\}$ *be a complete semi-inner product linear space with semi-inner product* J*,* $\mathcal{H}_1$ *a closed linear subspace of* $\mathcal{H}$ *and* $\mathcal{H}_0 = \{f \in \mathcal{H} : J(f) = 0\}$ *be the null space of* J *of finite dimension* l*, and basis* $\{\phi_i\}_{i=1}^l$*.* ***Note****: observe that by definition,* $\mathcal{H}_1$ *is RKHS with inner product* J*. Let* R_J *be the reproducing Kernel of* $\mathcal{H}_1$*,* $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathbb{X}\times\mathbb{R}$ *be a training set,* $\{\mathbf{z}_i\}_{i=1}^k \subset \mathbb{X}$*, and let* $\eta \in \mathcal{H}$ *have the form*

$$\eta(\mathbf{x}) = \sum_{i=1}^{l} d_i \phi_i(\mathbf{x}) + \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \mathbf{x})$$

then

$$\sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i))^2 + n\lambda J(\eta) = (\mathbf{y} - S\mathbf{d} - R\mathbf{c})^\intercal (\mathbf{y} - S\mathbf{d} - R\mathbf{c}) + n\lambda \mathbf{c}^\intercal Q \mathbf{c} \tag{B.2}$$

where $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ *with* $S_{i,j} = \phi_j(\mathbf{x}_i)$*,* $R \in \mathcal{M}_{n\times k}(\mathbb{R})$ *with* $R_{i,j} = R_J(\mathbf{x}_i, \mathbf{z}_j)$*,* $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ *with* $Q_{i,j} = R_J(\mathbf{z}_i, \mathbf{z}_j)$*,* $\mathbf{y} = (y_1 \cdots y_n)^\intercal \in \mathcal{M}_{n\times 1}(\mathbb{R})$*,* $\mathbf{d} = (d_1 \cdots d_l)^\intercal$ *and* $\mathbf{c} = (c_1 \cdots c_k)^\intercal$*.*

Proof.

With straightforward algebra one has that

$$\sum_{i=1}^{n} (y_i - \eta(\mathbf{x}_i))^2 = (\mathbf{y} - S\mathbf{d} - R\mathbf{c})^\intercal (\mathbf{y} - S\mathbf{d} - R\mathbf{c}),$$

while

$$\begin{aligned}
J(\eta) &= J(\eta, \eta) \\
&= J\left(\sum_{i=1}^{l} d_i\phi_i + \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \cdot), \sum_{i=1}^{l} d_i\phi_i + \sum_{i=1}^{k} c_i R_J(\mathbf{z}_i, \cdot)\right) \\
&= \sum_{i,j=1}^{l} d_i d_j J(\phi_i, \phi_j) + 2\sum_{i=1}^{k}\sum_{j=1}^{l} c_i d_j J(\phi_j, R_J(\mathbf{z}_i, \cdot)) + \sum_{i,j=1}^{k} d_i d_j J(R_J(\mathbf{z}_i, \cdot), R_J(\mathbf{z}_j, \cdot)) \\
&= \sum_{i,j=1}^{k} d_i J(R_J(\mathbf{z}_i, \cdot), R_J(\mathbf{z}_j, \cdot)) d_j \qquad \text{(B.3)} \\
&= \sum_{i,j=1}^{k} d_i R_J(\mathbf{z}_i, \mathbf{z}_j) d_j \qquad \text{(B.4)} \\
&= \mathbf{c}^\intercal Q \mathbf{c}.
\end{aligned}$$

where expression (B.3) was obtained using $J(\phi_i, f) \leq \sqrt{J(\phi_i)}\sqrt{J(f)} = 0$ by the Cauchy-Schwartz inequality and $J(\phi_i) = 0$. Equation (B.4) is obtained using the reproducing property. We have obtained (B.2). ■

Proposition 54

In the context of Theorem 5 or Proposition 6, the matrix $A(\lambda)$ such

$$\hat{\mathbf{y}} = A(\lambda)\mathbf{y}$$

has the representation

$$A(\lambda) = I_n - n\lambda\left(M^{-1} - M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right),$$

where $M = RQ^+R^\intercal + n\lambda I_n$

Proof. Using (3.4) or (3.5)

$$\begin{aligned}
\hat{\mathbf{y}} =& S\hat{\mathbf{d}} + R\hat{\mathbf{c}} \\
=& \left[S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1} + RQ^+R^\intercal\left(M^{-1} - M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\right]\mathbf{y} \\
=& \left[S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal + RQ^+R^\intercal - RQ^+R^\intercal S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal\right]M^{-1}\mathbf{y} \\
=& \left[\left\{I - RQ^+R^\intercal M^{-1}\right\}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal + RQ^+R^\intercal\right]M^{-1}\mathbf{y} \\
=& \left[\left\{M - RQ^+R^\intercal\right\}M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal + RQ^+R^\intercal\right]M^{-1}\mathbf{y} \\
=& \left[\left\{n\lambda I_n\right\}M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal + RQ^+R^\intercal\right]M^{-1}\mathbf{y} \\
=& \left[\left(RQ^+R^\intercal + n\lambda I_n\right)M^{-1} - n\lambda M^{-1} + n\lambda M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right]\mathbf{y} \\
=& \left[I_n - n\lambda\left(M^{-1} - M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\right]\mathbf{y}.
\end{aligned}$$

■

Corollary 55 *In the context of chapter 2, the smoothing matrix (2.21) can be written as*

$$A(\lambda) = I - n\lambda M^{-1}\left(I - S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)$$

where $M = Q + n\lambda I$.

Proof.

The context of the smoothing matrix (2.21) is the same as the Proposition 54 in the case that the knots and the training set are equal: $\{\mathbf{z}_i\}_{i=1}^k = \{\mathbf{x}_i\}_{i=1}^k$. Then $R = Q$, $M = Q + n\lambda I_n$ and the corollary follows by direct application of Proposition 54. ■

Theorem 56 (Existence and Uniqueness of Reproducing Kernel)

For every RKHS $\mathcal{H}$ of functions on $\mathbb{X}$, there corresponds a unique reproducing kernel $R(x, y)$, which is non-negative definite. Conversely, for every non-negative definite $R(x, y)$ on $\mathbb{X}$, there corresponds a unique RKHS $\mathcal{H}$ that has R as its reproducing kernel.

Proof.

The forward implication of the Theorem follows easily: if there were two reproducing kernels R_1 and R_2 in $\mathcal{H}$ then $\forall x, y \in \mathbb{X}$

$$R_1(x, y) = (R_1(x, \cdot), R_2(x, \cdot)) = R_2(x, y)$$

where the first equality is by the reproducing property of R_2 and the second equality is by the reproducing property of R_1.

We provide the idea of proof for the backward implication. The converse implication of the Theorem is proven by construction: take the space $\mathcal{H}^\star = \{f : f = \sum_{i=1} \alpha_i R(x_i, \cdot)\}$ defining the inner product in terms of R. Complete the space $H^\star$ by adding the limits of all the Cauchy sequences and define the norm of any limit of Cauchy sequences as the limit of the sequences of the norms; this definition of norm must be proven to be independent of the sequence converging to the same limit. Finally it has to be proven that the new reproducing kernel is the same as the original when restricted to $\mathcal{H}^\star$ Aronszajn (1950) , Gu (2013). ■

Theorem 57 *(Gu, 2013, Theorem 2.5, pag. 32)*

If the reproducing kernel R of a space $\mathcal{H}$ on a domain $\mathbb{X}$ can be decomposed into $R = R_0 + R_1$, where R_0 and R_1 are both non-negative definite, $\forall x \in \mathbb{X}$ $R_0(x, \cdot)$, $R_1(x, \cdot) \in \mathcal{H}$ and $(R_0(x, \cdot), R_1(y, \cdot)) = 0 \forall x, y \in \mathbb{X}$ then the RKHS's $\mathcal{H}_0$ and $\mathcal{H}_1$ corresponding to R_0 and R_1 form a tensor sum decomposition of $\mathcal{H}$. Conversely, if R_0 and R_1 are both non-negative definite and $\mathcal{H}_0 \cup \mathcal{H}_\infty = \{0\}$, then $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$ has a reproducing kernel $R = R_1 + R_2$.

Theorem 58 (Completness of Finite Normed Space)

Every finite dimensional normed space is complete.

Proof.

Let $\mathcal{H}$ be a normed vector space over $\mathbb{R}$ (or $\mathbb{C}$), with norm $\|\cdot\|$ and $\{\psi_i\}_{i=1}^N$ a basis for $\mathcal{H}$. For $\eta = \sum_{i=1}^N a_i\psi_i \in \mathcal{H}$ we can define another norm:

$$\|\eta\|^\star = \sqrt{\sum_{i=1}^N |a_i|^2}.$$

It would be needed to prove that $\|\cdot\|^\star$ is a norm but we let the reader to prove it since it is straight forward. Since all norms are equivalent in a finite dimensional vector space, there exist a constant $c > 0$ such for all $\eta \in \mathcal{H}$

$$\frac{1}{c}\|\eta\|^\star \leq \|\eta\| \leq c\|\eta\|^\star$$

Let $\{\eta_i\}_{i=1}^N$ be a Cauchy sequence in $(\mathcal{H}, \|\cdot\|)$, and write $\eta_i = \sum_{j=1}^N a_j^{(i)}\psi_j$. Therefore, $\forall\epsilon,\ \exists M \in \mathbb{N}$ such that $\forall n, m > M$

$$\epsilon > \|\eta_n - \eta_m\| \geq \frac{1}{c}\|\eta_n - \eta_m\|^\star = \frac{1}{c}\sqrt{\sum_{i=1}^N \left|a_i^{(n)} - a_i^{(m)}\right|^2} \geq \frac{1}{c}\left|a_i^{(n)} - a_i^{(m)}\right|$$

which show that $\{a_1^j\}_{j=1}^\infty, \cdots, \{a_N^j\}_{j=1}^\infty$ are Cauchy sequences in $\mathbb{R}$ (or $\mathbb{C}$). Since $\mathbb{R}$ (or $\mathbb{C}$) is a complete space there exists $\{a_i\}_{i=1}^N \subset \mathbb{R}\,(\mathbb{C})$ such $\lim_{j\to\infty} a_i^{(j)} = a_i$, $i \in \{1, \cdots, N\}$.

Let $\eta = \sum_{i=1}^N a_i\psi_i \in \mathcal{H}$ and $\epsilon > 0$, then

$$|\eta - \eta_i| \leq c|\eta - \eta_i|^\star = c\sqrt{\sum_{j=1}^N \left|a_j - a_j^{(i)}\right|^2}, \tag{B.5}$$

since $\lim_{j\to\infty} a_i^{(j)} = a_i$, $i \in \{1, \cdots, N\}$ we can take $M \in \mathbb{N}$ large enough such that if $i > M$, $\sqrt{\sum_{j=1}^N \left|a_j - a_j^{(i)}\right|^2} < c\epsilon$; by (B.5) we conclude $|\eta - \eta_i| \leq \epsilon$, or $\lim_{i\to\infty}\eta_i = \eta$ in $(\mathcal{H}, \|\cdot\|)$. Then $(\mathcal{H}, \|\cdot\|)$ is a complete space. ■

Proposition 59

Let $\mathcal{H}$ be the space of polynomials with d indeterminates and degree smaller than m. Let $\{\psi_i\}_{i=1}^l$ be a basis of $\mathcal{H}$, let $N \in \mathbb{N}$, $\{u_i\}_{i=1}^N \subset \mathbb{R}^d$, $\{p_i\}_{i=1}^N \subset \mathbb{R}$, $p_i > 0$ and $\sum_{i=1}^N p_i = 1$ define

$$(\eta, \zeta)_0 = \sum_{i=1}^N p_i\eta(u_i)\zeta(u_i) \tag{B.6}$$

where $\{u_i\}_{i=1}^N$ *and* $\{p_i\}_{i=1}^N$ *are specified such that the matrix with* $(i,j)th$ *entry* $(\psi_i, \psi_j)_0$ *is non-singular. Then*

i) $\}\mathcal{H}, (\cdot,\cdot)_0\}$ *is a complete inner product space.*

ii) *Let* $\{\phi_i\}_{i=1}^l \subset \mathcal{H}$ *a orthonormal basis. Then*

$$R_0(x,y) = \sum_{i=1}^{l} \phi_i(x)\phi_i(y) \tag{B.7}$$

is a reproducing kernel in $\mathcal{H}$*.*

iii) $\}\mathcal{H}, (\cdot,\cdot)_0\}$ *is a RKHS with* R_0 *as its reproducing kernel.*

Proof.

i) We'll prove $(\mathcal{H}, (\cdot,\cdot)_0)$ is a complete inner product space.

- The properties that make $\mathcal{H}$ a vector space over $\mathbb{R}$ with the operation sum of polynomials and multiplication by scalars are direct to prove.
- The properties of symmetry and linearity with the operations of sum and multiplication by scalars are straight forward to prove from the definition of $(\cdot,\cdot)_0$.
- We show now $(\cdot,\cdot)_0$ is positive definite. Let $\eta = \sum_{i=1}^l a_i\psi_i$. Then,

$$\begin{aligned}
(\eta,\eta)_0 &= \left(\sum_{i=1}^{l} a_i\psi_i, \sum_{i=1}^{l} a_i\psi_i\right)_0 \\
&= \sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j \left(\psi_i, \psi_j\right)_0 \\
&= \sum_{i=1}^{l}\sum_{j=1}^{l} a_i a_j \sum_{k=1}^{N} p_k \psi_i\left(u_k\right)\psi_j\left(u_k\right) \\
&= \begin{pmatrix} a_1 \\ \vdots \\ a_l \end{pmatrix}^{\intercal} \begin{pmatrix} \sum_{k=1}^N p_k\psi_1\left(u_k\right)\psi_1\left(u_k\right) & \cdots & \sum_{k=1}^N p_k\psi_1\left(u_k\right)\psi_l\left(u_k\right) \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^N p_k\psi_l\left(u_k\right)\psi_1\left(u_k\right) & \cdots & \sum_{k=1}^N p_k\psi_l\left(u_k\right)\psi_l\left(u_k\right) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_l \end{pmatrix} \\
&= \alpha^{\intercal}\boldsymbol{\Psi}^{\intercal}\boldsymbol{\Psi}\alpha
\end{aligned} \tag{B.8}$$

$$= (\mathbf{\Psi}\alpha)^{\intercal} (\mathbf{\Psi}\alpha)$$

$$= \|\mathbf{\Psi}\alpha\|^2_{euclidean}$$

$$\geq 0,$$

where

$$\mathbf{\Psi} = \begin{pmatrix} \sqrt{p_1}\psi_1(u_1) & \sqrt{p_1}\psi_2(u_1) & \cdots & \sqrt{p_1}\psi_l(u_1) \\ \sqrt{p_2}\psi_1(u_2) & \sqrt{p_2}\psi_2(u_2) & \cdots & \sqrt{p_2}\psi_l(u_2) \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{p_N}\psi_1(u_N) & \sqrt{p_N}\psi_2(u_N) & \cdots & \sqrt{p_N}\psi_l(u_N) \end{pmatrix} \text{ and } \alpha = \begin{pmatrix} a_1 \\ \vdots \\ a_l \end{pmatrix}$$

hence $(\cdot,\cdot)_0$ is non-negative definite.

By hypothesis, $\{u_i\}_{i=1}^N$ and $\{p_i\}_{i=1}^N$ were specified with the condition that the matrix in (*B*.8): $\mathbf{\Psi}^{\intercal}\mathbf{\Psi}$ is non-singular, therefore $(\eta,\eta)_0 = 0$ if and only if $\alpha = \mathbf{0}$ or equivalently if $\eta = 0$. We have proven that $(\mathcal{H}, (\cdot,\cdot)_0)$ is a inner space.

- It is proven that $\{\mathcal{H}, (\cdot,\cdot)_0\}$ is a complete inner product space using that it is of finite dimension and Theorem 58 concludes the proof.

ii) We will now show that R_0 is a reproducing kernel in $\mathcal{H}$.

- The symmetry of R_0 is trivial.
- We prove now R_0 is non negative definite. Let $\{x_i\}_{i=1}^n \subset \mathbb{X}$ and $\{a_i\}_{i=1}^n \subset \mathbb{X}$ for $n \in \mathbb{N}$. Then,

$$\sum_{i,j=1}^{n} a_i a_j R_0(x_i, x_j) = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}^{\intercal} \begin{pmatrix} R_0(x_1,x_1) & \cdots & R_0(x_1,x_n) \\ \vdots & \ddots & \vdots \\ R_0(x_n,x_1) & \cdots & R_0(x_n,x_n) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

$$= \alpha^{\intercal}\mathbf{\Phi}^{\intercal}\mathbf{\Phi}\alpha$$

$$= (\mathbf{\Phi}\alpha)^{\intercal} (\mathbf{\Phi}\alpha)$$

$$= \|\mathbf{\Phi}\alpha\|^2_{euclidean}$$

$$\geq 0.$$

where

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1(x_1) & \phi_1(x_2) & \cdots & \phi_1(x_n) \\ \phi_2(x_1) & \phi_2(x_2) & \cdots & \phi_2(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_l(x_1) & \phi_l(x_2) & \cdots & \phi_l(x_n) \end{pmatrix} \text{ and } \alpha = \begin{pmatrix} a_1 \\ \vdots \\ a_l \end{pmatrix}.$$

Hence R_0 is non-negative definite.

- The reproducing property follows as

$$\begin{aligned} (R_0(x,\cdot), f)_0 &= \left(\sum_{i=1}^{l} \phi_i(x)\phi_i, \sum_{i=1}^{l} a_i\phi_i \right)_0 \\ &= \sum_{i=1}^{l} \phi_i(x) \sum_{j=1}^{l} a_j \left(\phi_i, \phi_j\right)_0 \\ &= \sum_{i=1}^{l} \phi_i(x) a_i \\ &= f(x), \end{aligned}$$

where it was used that any $f \in \mathcal{N}_J$ can be written as $f = \sum_{i=1}^{l} a_i\phi_i$ for some a_is and that $\{\phi_i\}_{i=1}^{l}$ is orthonormal with respect to the inner product $(\cdot,\cdot)_0$.

Hence R_0 is a reproducing kernel in $\mathcal{H}$.

iii) In order to prove that $\mathcal{H}$, it is a *RKHS* is only needed to show that for any $x \in \mathbb{X}$, the evaluation functional $[x]\eta := \eta(x)$ is continuous in $\{\mathcal{H}, (\cdot,\cdot)_0\}$.

Let $\eta \in \mathcal{H}$, $x \in \mathbb{X}$ and $\{\eta_i\}_{i=1}^{\infty} \subset \mathcal{H}$ a sequence of functions converging to η. Then we have:

$$|\eta(x) - \eta_i(x)| = |(R_0(x), \eta - \eta_i)| \leq \|R_0(x,\cdot)\| \|\eta - \eta_i\|_0 = \sqrt{R_0(x,x)} \|\eta - \eta_i\|_0 \qquad \text{(B.9)}$$

where the first and last equality are by the reproducing property of R_0 and the inequality is by the Cauchy-Schwartz inequality. Therefore the convergence of η_i into η implies the convergence of $\eta_i(x)$ into $\eta(x)$ and thus the projection functional $[x]$ is continuous. We have proven that $\mathcal{H}$ is a *RKHS* and by Theorem 56 its reproducing kernel is R_0.

R_0 is symmetric, non-negative definite (because $R_0(x_i, y_j)$ is the Gram matrix of linear independent vectors Bhatia (2009)) and has the reproducing property. ■

Proposition 60 (Projection onto $\mathcal{N}_{J_m^d}$)

In the context of Section 2.1.1, the projection of $f \in \mathcal{H}$ onto $\mathcal{N}_J$ is defined by the operator P through

$$(Pf)(x) = \sum_{\nu=1}^{l} (f, \phi_\nu)_0 \, \phi_\nu(x).$$

Proof. Direct proof applying the expression. Tedious procedure. ■

Proposition 61 (Reproducing Kernel of $\mathcal{H} \ominus \mathcal{N}_J$) *(Wahba and Wendelberger (1980))*

In the context of section 2.1.1.1 the bi-linear form

$$R_1(\mathbf{x}, \mathbf{y}) = \left(I - P_{(\mathbf{x})}\right)\left(I - P_{(\mathbf{y})}\right) E\left(\|\mathbf{x} - \mathbf{y}\|\right)$$

is the reproducing kernel of $\mathcal{H} \ominus \mathcal{N}_J$.

Proposition 62 (Orthonormal Basis in $\mathcal{N}_{J_m^d}$) *(Gu, 2013, Exercise 4.14, pag 169)*

Let $\{\psi_i\}_{i=1}^l$ be a set of polynomials consisting of d variables that span $\mathcal{N}_{J_m^d}$, $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{X}$ and $\tilde{S} \in \mathcal{M}_{N\times l}(\mathbb{R})$ such $\tilde{S}_{i,j} = \psi_j(\mathbf{x}_i)$. Let $\tilde{S} = F_1 U$ denote the QR-decomposition of $\tilde{S}$, then the polynomials $\{\phi_i\}_{i=1}^l$ defined as $\phi_i = \sqrt{n}\sum_{j=1}^l (U^{-\intercal})_{i,j}\,\psi_j$ form an a orthonormal basis in $\mathcal{N}_{J_m^d}$ with respect to the inner product $(f,g)_0 = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)g(\mathbf{x}_i)$, and $\sqrt{n}(F_1)_{i,j} = \phi_j(\mathbf{x}_i)$.

Proof.

The orthonormal property of the basis $\{\phi_i\}_{i=1}^l$ is directly observed by expanding $(\phi_i, \phi_j)_0$ from one side, and on the other side using that F_1 has orthonormal columns -because $\tilde{S}$ is full column rank- as measured by the usual euclidean inner product in $\mathbb{R}^d$ and that U is upper triangular. Finally checking that both expansions lead to the same expression. ■

Theorem 63 (Construction Reproducing Kernel of Tensor Product of *RKHS*)

For $R_{(1)}(x_{(1)}, y_{(1)})$ non-negative definite on $\mathbb{X}_{(1)}$ and $R_{(2)}(x_{(2)}, y_{(2)})$ non-negative definite on $\mathbb{X}_{(2)}$ then $R(x,y) = R_{(1)}(x_{(1)}, y_{(1)})R_{(2)}(x_{(2)}, y_{(2)})$ is non-negative on $\mathbb{X} = \mathbb{X}_{(1)} \times \mathbb{X}_{(2)}$.

Proof. (Aronszajn, 1950, pag. 358). ■

Proposition 64 (Gradient of Minimization Objective)

Let $S \in \mathcal{M}_{n\times l}(\mathbb{R})$, $R \in \mathcal{M}_{n\times k}(\mathbb{R})$, $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$, $\mathbf{y} \in \mathcal{M}_{n\times 1}(\mathbb{R})$, $\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right) \in \mathcal{M}_{(l+k)\times 1}(\mathbb{R})$. *Define the function*

$$\mathcal{L}\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\right] = (\mathbf{y} - S\mathbf{d} - R\mathbf{c})^{\intercal}(\mathbf{y} - S\mathbf{d} - R\mathbf{c}) + n\lambda \mathbf{c}^{\intercal} Q \mathbf{c}$$

The gradient of $\mathcal{L}$ *is*

$$\nabla\mathcal{L}\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\right] = 2\left[\begin{pmatrix} S^{\intercal}S & S^{\intercal}R \\ R^{\intercal}S & R^{\intercal}R + n\lambda Q^{\intercal}\end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} - \begin{pmatrix} S^{\intercal}\mathbf{y} \\ R^{\intercal}\mathbf{y}\end{pmatrix}\right]^{\intercal} \tag{B.10}$$

and $\nabla\mathcal{L}\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\right] = \mathbf{0}$ *if and only if*

$$\begin{aligned} S^{\intercal}\{S\mathbf{d} + R\mathbf{c} - \mathbf{y}\} &= \mathbf{0} \\ R^{\intercal}\left\{S\mathbf{d} + \left(R + n\lambda\left(R^{\intercal}\right)^{-1}_{Right} Q\right)\mathbf{c} - \mathbf{y}\right\} &= \mathbf{0} \end{aligned} \tag{B.11}$$

where for a matrix $A \in \mathcal{M}_{\alpha,\beta}(\mathbb{R}), \alpha < \beta$ *of full rank,* $(A)^{-1}_{Right} = A^{\intercal}\left(AA^{\intercal}\right)^{-1}$ *is the right inverse of* A*.*

Proof.

First we re-write $\mathcal{L}$ as:

$$\begin{aligned}\mathcal{L}\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\right] = \mathbf{y}^{\intercal}\mathbf{y} &- \begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}^{\intercal}\begin{pmatrix} S & R\end{pmatrix}^{\intercal}\mathbf{y} - \mathbf{y}^{\intercal}\begin{pmatrix} S & R\end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} \\ &+ \begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}^{\intercal}\begin{pmatrix} S & R\end{pmatrix}^{\intercal}\begin{pmatrix} S & R\end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} + n\lambda\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}^{\intercal}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & Q\end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}\end{aligned}$$

We proceed directly taking the derivatives from previous expression:

$$\nabla\mathcal{L}\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\right] = -2\mathbf{y}^{\intercal}\begin{pmatrix} S & R\end{pmatrix} + 2\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}^{\intercal}\begin{pmatrix} S^{\intercal}S & S^{\intercal}R \\ R^{\intercal}S & R^{\intercal}R\end{pmatrix} + 2n\lambda\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}^{\intercal}\begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & Q\end{pmatrix}$$

$$\left(\nabla\mathcal{L}\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\right]\right)^{\intercal} = 2\left[\begin{pmatrix} S^{\intercal}S & S^{\intercal}R \\ R^{\intercal}S & R^{\intercal}R + n\lambda Q^{\intercal}\end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} - \begin{pmatrix} S^{\intercal}\mathbf{y} \\ R^{\intercal}\mathbf{y}\end{pmatrix}\right]$$

we have obtained (*B*.10). Expressions (*B*.11) can be obtained equating the gradient to $\mathbf{0}$ and simple algebra leads to the result. ■

Lemma 65 *For $n, S \in \mathbb{N}$, the total number of solution for $\{a_i\}_{i=1}^n \subset \mathbb{N}$ to the equation $\sum_{i=1}^n a_i = S$ is $\binom{S-1}{n-1}$.*

Proof.

Consider S balls in a line. We would like to partition these S balls in n parts by drawing sticks between these balls. The number of balls in the $i'th$ partition represents a_i and we will have $\sum_{i=1}^n a_i = S$ with $\{a_i\}_{i=1}^n \subset \mathbb{N}$.

But there are $S-1$ gaps between these S balls so the total number of ways of drawing $n-1$ sticks between the balls is $\binom{S-1}{n-1}$.

Therefore the total number of solution for $\{a_i\}_{i=1}^n \subset \mathbb{N}$ to the equation $\sum_{i=1}^n a_i = S$ is $\binom{S-1}{n-1}$.

■

Lemma 66 *For $n, S \in \mathbb{N}$ the total number of solutions for $\{b_i\}_{i=1}^n \subset \mathbb{N} \cup \{0\}$ to the equation $\sum_{i=1}^n b_i = S$ is $\binom{S+n-1}{n-1}$.*

Proof.

Let us write $b_i = a_i - 1$ then we have the equivalent problem of finding the number of solutions for $\{a_i\}_{i=1}^n \subset \mathbb{N}$ to the equation $\sum_{i=1}^n (a_i - 1) = S$ or $\sum_{i=1}^n a_i = S + n$.

By Lemma 65 the number of solutions is $\binom{S+n-1}{n-1}$. ■

Lemma 67 *The null space of J_m^d has dimension $\binom{d+m-1}{d}$.*

Proof.

Observe that $J_m^d(f) = 0$ if and only if $\dfrac{\partial^m f}{\partial X_1^{\alpha_1} \cdots \partial X_d^{\alpha_d}}(x) = 0$ for $\alpha_1 + \cdots + \alpha_d = m$, and $\alpha_i \in \mathbb{N} \cup \{0\}$.

Using the Taylor expansion Corwin (1982), Stewart (2011), Königsberger (2013) can be seen that the null space of J_m^d consist of polynomials of up to $m-1$ total order since all but the first $k-1$ terms on the Taylor expansion must be 0 leaving only polynomials of degree smaller than m.

A basis for the null space is then $\{1^{\alpha_0} x_1^{\alpha_1} x_d^{\alpha_d}\}_{\alpha_0+\alpha_1+\cdots+\alpha_d=m-1}$, therefore the size of the basis is the number of solutions for $\{\alpha_i\}_{i=0}^d \subset \mathbb{N} \cup \{0\}$ to the equation $\sum_{i=0}^d \alpha_i = m - 1$.

Taking $\alpha_i = b_{i-1}$ we have the equation $\sum_{i=1}^{d+1} b_i = m-1$. By Lemma 66 there are $\binom{(m-1)+(d+1)-1}{(d+1)-1} = \binom{m+d-1}{d}$ solutions. Therefore the size of the basis $\{1^{\alpha_0} x_1^{\alpha_1} x_d^{\alpha_d}\}_{\alpha_0+\alpha_1+\cdots+\alpha_d=m-1}$ is $\binom{(m-1)+(d+1)-1}{(d+1)-1} = \binom{m+d-1}{d}$ as well as the dimension of the null space of J_m^d. ■

Lemma 68 (Duchon (1977))

The quadratic functional J_m^d defined in (2.24) is a square semi norm.

Lemma 69 (Duchon (1977))

The quadratic functional J_m^d defined in (2.24) is invariant under coordinate rotations.

Lemma 70 *Let $\{\phi\}_{i=1}^M$ be a collection of functions with domain $\mathbb{X} \neq \emptyset$ and range in $\mathbb{R}$. Define $\eta(\mathbf{x}) := \sum_{i=1}^M d_i \phi_i(\mathbf{x})$. If $\{d_i\}_{i=1}^M \overset{iid}{\sim} N(0, \tau^2)$ then η is a Gaussian Process with mean 0 and $\mathbb{E}(\eta(\mathbf{x})\eta(\mathbf{y})) = \tau^2 \sum_{i=1}^M \phi_i(\mathbf{x})\phi_i(\mathbf{y})$.*

Proof.

First we prove that η is indeed a Gaussian process by showing that $\sum_{i=1}^n a_i \eta(\mathbf{x}_i)$, $\{a_i\}_{i=1}^n \subset \mathbb{R}$ has normal distribution:

$$\begin{aligned}
\sum_{i=1}^n a_i \eta(\mathbf{x}_i) &= \sum_{i=1}^n a_i \sum_{j=1}^M d_j \phi_j(\mathbf{x}_i) \\
&= \sum_{j=1}^M d_j \left(\sum_{i=1}^n a_i \phi_j(\mathbf{x}_i) \right) \\
&\sim N\left(0, \tau^2 \sum_{j=1}^M \left(\sum_{i=1}^n a_i \phi_j(\mathbf{x}_i) \right)^2 \right).
\end{aligned}$$

Therefore η is a Gaussian process.

We have as well that $\mathbb{E}(\eta(\mathbf{x})) = \sum_{i=1}^M \mathbb{E}(d_i)\phi_i(\mathbf{x}) = 0$ and

$$\begin{aligned}
\mathbb{E}(\eta(\mathbf{x})\eta(\mathbf{y})) &= \mathbb{E}\left(\sum_{i=1}^M d_i^2 \phi_i(\mathbf{x})\phi_i(\mathbf{y}) + \sum_{i \neq j}^M d_i d_j \phi_i(\mathbf{x})\phi_j(\mathbf{y}) \right) \\
&= \mathbb{E}\left(\sum_{i=1}^M d_i^2 \phi_i(\mathbf{x})\phi_i(\mathbf{y}) \right) + \mathbb{E}\left(\sum_{i \neq j}^M d_i d_j \phi_i(\mathbf{x})\phi_j(\mathbf{y}) \right) \\
&= \sum_{i=1}^M \mathbb{E}\left(d_i^2\right) \phi_i(\mathbf{x})\phi_i(\mathbf{y})
\end{aligned}$$

$$= \sum_{i=1}^{M} \tau^2 \phi_i(\mathbf{x})\phi_i(\mathbf{y}).$$

■

Lemma 71 *Let $\{\psi\}_{i=1}^n$ be a collection of functions with domain $\mathbb{X} \neq \emptyset$ and range in $\mathbb{R}$. Let $\eta(\mathbf{x}) := \sum_{i=1}^n c_i\psi_i(\mathbf{x})$. If $(c_1 \dots c_n)^\intercal \sim N_n(\mathbf{0}, \mathbf{\Sigma})$ then η is a Gaussian Process with mean 0 and*

$$\mathbb{E}\left(\eta(\mathbf{x})\eta(\mathbf{y})\right) = \left(\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x})\right) \mathbf{\Sigma} \left(\psi_1(\mathbf{y}) \dots \psi_n(\mathbf{y})\right)^\intercal.$$

Proof.

Let $\left(\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x})\right)^\intercal = \mathbf{\Sigma}^{\frac{1}{2}} \left(\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x})\right)^\intercal$ where $\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}} = \mathbf{\Sigma}$. We have that

$$\begin{aligned}
\eta(\mathbf{x}) &= \sum_{i=1}^{n} c_i\psi_i(\mathbf{x}) \\
&= \left(c_1 \dots c_n\right) \left(\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x})\right)^\intercal \\
&= \left(c_1 \dots c_n\right) \mathbf{\Sigma}^{-\frac{1}{2}} \left(\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x})\right)^\intercal \\
&= \left(\mathbf{\Sigma}^{-\frac{1}{2}} \left(c_1 \dots c_n\right)^\intercal\right)^\intercal \left(\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x})\right)^\intercal \\
&= \left(d_1 \dots d_n\right) \left(\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x})\right)^\intercal \\
&= \sum_{i=1}^{n} d_i\phi_i(\mathbf{x}),
\end{aligned}$$

where $\left(d_1 \dots d_n\right)^\intercal = \mathbf{\Sigma}^{-\frac{1}{2}} \left(c_1 \dots c_n\right)^\intercal$ and thus $\left(d_1 \dots d_n\right)^\intercal \sim N_n\left(\mathbf{0}, \mathbf{I}_n\right)$. Then by Lemma 70, η is a Gaussian process, with mean 0 and

$$\begin{aligned}
\mathbb{E}(\eta(\mathbf{x})\eta(\mathbf{y})) &= \sum_{i=1}^{n} \phi_i(\mathbf{x})\phi_i(\mathbf{y}) \\
&= \left(\phi_1(\mathbf{x}) \dots \phi_n(\mathbf{x})\right) \left(\phi_1(\mathbf{y}) \dots \phi_n(\mathbf{y})\right)^\intercal \\
&= \left(\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x})\right) \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{\Sigma}^{\frac{1}{2}} \left(\psi_1(\mathbf{y}) \dots \psi_n(\mathbf{y})\right)^\intercal \\
&= \left(\psi_1(\mathbf{x}) \dots \psi_n(\mathbf{x})\right) \mathbf{\Sigma} \left(\psi_1(\mathbf{y}) \dots \psi_n(\mathbf{y})\right)^\intercal.
\end{aligned}$$

■

Proposition 72 *[Full Conditional Posterior of Coefficients* $\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)$*. Version 1.]*

For the observed pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ *consider the model with* $\sigma^2 > 0$ *known given by*

$$\begin{aligned} y_i &= \sum_{j=1}^{k} d_j \phi_j(\mathbf{x}_i) + \sum_{j=1}^{l} d_j \psi_j(\mathbf{x}_i) + \epsilon_i, \\ \epsilon_i &\overset{iid}{\sim} N_1(0, \sigma^2), \end{aligned}$$

where $\{\phi\}_{i=1}^k$ *and* $\{\psi\}_{i=1}^l$ *are known functions. For* $b > 0, \tau > 0$ *and* Q *a* $l \times l$ *positive definite matrix, consider the priors*

$$\begin{aligned} d_i &\overset{iid}{\sim} N_1(0, \tau^2) \\ (c_1\, c_2\, \cdots\, c_l)^\intercal &\sim N_l(\mathbf{0}, b\,Q) \\ (d_1\, d_2\, \cdots\, d_k)^\intercal &\perp (c_1\, c_2\, \cdots\, c_l)^\intercal \\ (d_1\, \cdots\, d_k\, c_1\, \cdots\, c_l)^\intercal &\perp (\epsilon_1 \cdots \epsilon_n)^\intercal. \end{aligned}$$

The posterior of $\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)$ *is* $N_{k+l}\left(\mu_{\mathbf{dc}_\rho}, b\boldsymbol{\Sigma}_{\mathbf{dc}_\rho}\right)$ *where*

$$\mu_{\mathbf{dc}_\rho} = \begin{pmatrix} \rho S^\intercal (\rho S S^\intercal + M)^{-1} \\ Q R^\intercal (\rho S S^\intercal + M)^{-1} \end{pmatrix} \mathbf{y} \tag{B.12}$$

$$\boldsymbol{\Sigma}_{\mathbf{dc}_\rho} = \begin{pmatrix} \left(\rho I_k - \rho S^\intercal (\rho S S^\intercal + M)^{-1} \rho S\right) & \left(-\rho S^\intercal (\rho S S^\intercal + M)^{-1} R Q\right) \\ \left(-Q R^\intercal (\rho S S^\intercal + M)^{-1} \rho S\right) & \left(Q - Q R^\intercal (\rho S S^\intercal + M)^{-1} R Q\right) \end{pmatrix}. \tag{B.13}$$

and S *is a* $n \times l$ *matrix with entry* $S_{i,j} = \phi_j(\mathbf{x}_i)$*,* R *is a* $n \times l$ *matrix with entries* $R_{i,j} = \psi_j(\mathbf{x}_i)$*,* $M = RQR^\intercal + n\lambda I_n$*,* $n\lambda = \frac{\sigma^2}{b}$ *and* $\rho = \frac{\tau^2}{b}$*.*

Proof.

Observe that

$$\begin{aligned} \mathbb{E}(\mathbf{d}) &= \mathbf{0}_k, \\ \mathbb{E}(\mathbf{c}) &= \mathbf{0}_l, \\ \mathbb{E}(\mathbf{y}) &= \mathbb{E}(S\mathbf{d} + R\mathbf{c} + \epsilon) = S\,\mathbb{E}(\mathbf{d}) + R\,\mathbb{E}(\mathbf{c}) + \mathbb{E}(\epsilon) = \mathbf{0}_n, \\ Var(\mathbf{y}) &= Var(S\mathbf{d} + R\mathbf{c} + \epsilon) = S Var(\mathbf{d}) S^\intercal + R Var(\mathbf{c}) R^\intercal + Var(\epsilon) \\ &= \tau^2 S S^\intercal + b R Q R^\intercal + \sigma^2 I_n, \end{aligned}$$

$$Var\left[\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}\right] = \begin{pmatrix}\tau^2 I_k & \mathbf{0}_{k\times l}\\ \mathbf{0}_{l\times k} & bQ\end{pmatrix},$$

$$\begin{aligned}
Cov(y_i, d_j) &= Cov\left(\sum_{\nu=1}^{k} d_\nu \phi_\nu(\mathbf{x}_i) + \sum_{\nu=1}^{l} d_\nu \psi_\nu(\mathbf{x}_i) + \epsilon_i, d_j\right)\\
&= \phi_j(\mathbf{x}_i)\, Cov(d_j, d_j) = \tau^2 \phi_j(\mathbf{x}_i) = \tau^2 S_{i,j},\\
Cov(y_i, c_j) &= Cov\left(\sum_{\nu=1}^{k} d_\nu \phi_\nu(\mathbf{x}_i) + \sum_{\nu=1}^{l} c_\nu \psi_\nu(\mathbf{x}_i) + \epsilon_i, c_j\right)\\
&= \sum_{\nu=1}^{l} \psi_\nu(\mathbf{x}_i)\, Cov(c_\nu, c_j) = b\sum_{\nu=1}^{l} R_{i,\nu} Q_{\nu,j} = b(RQ)_{i,j},
\end{aligned}$$

$$\pi\left[\begin{pmatrix}\mathbf{y}\\ \mathbf{d}\\ \mathbf{c}\end{pmatrix}\right] = \pi\left[\mathbf{y} \mid \begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}\right] \pi\left[\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix}\right],$$

then $\pi\left[\begin{pmatrix}\mathbf{y}\\ \mathbf{d}\\ \mathbf{c}\end{pmatrix}\right] = N_{n+l+k}(\mathbf{0}, \mathbf{\Sigma}_{ydc})$ where

$$\mathbf{\Sigma}_{ydc} = \begin{pmatrix} \left(\tau^2 SS^\intercal + bRQR^\intercal + \sigma^2 I_n\right) & \begin{pmatrix}\tau^2 S & bRQ\end{pmatrix} \\ \begin{pmatrix}\tau^2 S^\intercal\\ bQR^\intercal\end{pmatrix} & \begin{pmatrix}\tau^2 I_k & \mathbf{0}\\ \mathbf{0} & bQ\end{pmatrix} \end{pmatrix}.$$

By a known result in multivariate statistics for multivariate normal distributions (Bilodeau and Brenner (2008)), we obtain that the posterior distribution $\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right) \mid \mathbf{y}\right]$ is normal with mean and covariance described below:

$$\begin{aligned}
\mathbb{E}\left[\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} \mid \mathbf{y}\right] &= \begin{pmatrix}\tau^2 S^\intercal\\ bQR^\intercal\end{pmatrix}\left(\tau^2 SS^\intercal + bRQR^\intercal + \sigma^2 I_n\right)^{-1}\mathbf{y}\\
&= \begin{pmatrix}\frac{\tau^2}{b} S^\intercal\\ QR^\intercal\end{pmatrix}\left(\frac{\tau^2}{b} SS^\intercal + RQR^\intercal + \frac{\sigma^2}{b} I_n\right)^{-1}\mathbf{y}\\
&= \begin{pmatrix}\rho S^\intercal\\ QR^\intercal\end{pmatrix}\left(\rho SS^\intercal + M\right)^{-1}\mathbf{y}
\end{aligned}$$

$$= \begin{pmatrix} \rho S^\intercal \left(\rho S S^\intercal + M\right)^{-1} \\ Q R^\intercal \left(\rho S S^\intercal + M\right)^{-1} \end{pmatrix} \mathbf{y},$$

and

$$\begin{aligned} Var\left[\begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} \mid \mathbf{y}\right] =& \begin{pmatrix} \tau^2 I_k & \mathbf{0} \\ \mathbf{0} & bQ \end{pmatrix} \\ & - \begin{pmatrix} \tau^2 S^\intercal \\ bQR^\intercal \end{pmatrix} \left(\tau^2 S S^\intercal + bRQR^\intercal + \sigma^2 I_n\right)^{-1} \begin{pmatrix} \tau^2 S & bRQ \end{pmatrix} \\ =& b \begin{pmatrix} \frac{\tau^2}{b} I_k & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} - b \begin{pmatrix} \frac{\tau^2}{b} S^\intercal \\ QR^\intercal \end{pmatrix} \left(\frac{\tau^2}{b} S S^\intercal + M\right)^{-1} \begin{pmatrix} \frac{\tau^2}{b} S & RQ \end{pmatrix} \\ =& b \begin{pmatrix} \rho I_k & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} - b \begin{pmatrix} \rho S^\intercal \\ QR^\intercal \end{pmatrix} \left(\rho S S^\intercal + M\right)^{-1} \begin{pmatrix} \rho S & RQ \end{pmatrix} \\ =& b \left[\begin{pmatrix} \rho I_k & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} - \begin{pmatrix} \rho S^\intercal \left(\rho S S^\intercal + M\right)^{-1} \begin{pmatrix} \rho S & RQ \end{pmatrix} \\ QR^\intercal \left(\rho S S^\intercal + M\right)^{-1} \begin{pmatrix} \rho S & RQ \end{pmatrix} \end{pmatrix} \right] \\ =& b \left\{ \begin{pmatrix} \rho I_k & \mathbf{0} \\ \mathbf{0} & Q \end{pmatrix} - \begin{pmatrix} \left[\rho S^\intercal \left(\rho S S^\intercal + M\right)^{-1} \rho S\right] & \left[\rho S^\intercal \left(\rho S S^\intercal + M\right)^{-1} RQ\right] \\ \left[QR^\intercal \left(\rho S S^\intercal + M\right)^{-1} \rho S\right] & \left[QR^\intercal \left(\rho S S^\intercal + M\right)^{-1} RQ\right] \end{pmatrix} \right\} \\ =& b \begin{pmatrix} \left[\rho I_k - \rho S^\intercal \left(\rho S S^\intercal + M\right)^{-1} \rho S\right] & \left[-\rho S^\intercal \left(\rho S S^\intercal + M\right)^{-1} RQ\right] \\ \left[-QR^\intercal \left(\rho S S^\intercal + M\right)^{-1} \rho S\right] & \left[Q - QR^\intercal \left(\rho S S^\intercal + M\right)^{-1} RQ\right] \end{pmatrix}. \end{aligned}$$

■

Lemma 73

Suppose M is a symmetric and nonsingular matrix and S is a full column rank matrix, then

$$\lim_{\rho \to \infty} \left(\rho S S^\intercal + M\right)^{-1} = M^{-1} - M^{-1} S \left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1}$$

$$\lim_{\rho \to \infty} \rho S^\intercal \left(\rho S S^\intercal + M\right)^{-1} = \left(S^\intercal M^{-1} S\right)^{-1} S^\intercal M^{-1}$$

$$\lim_{\rho \to \infty} \rho I - \rho^2 S^\intercal \left(\rho S S^\intercal + M\right)^{-1} S^\intercal = \left(S^\intercal M^{-1} S\right)^{-1}.$$

Proof.

We describe the idea of the proof; the details can be found in (Wahba, 1978, p. 367) and

(Wahba, 1983, p. 137). The proof is based on the identity

$$(\rho SS^\intercal + M)^{-1} = M^{-1} - M^{-1}S\left(S^\intercal M^{-1}S\right)^{-1}\left(I + \rho^{-1}\left(S^\intercal M^{-1}S\right)^{-1}\right)^{-1}S^\intercal M^{-1} \tag{B.14}$$

which can be proven directly. Using (B.14) with some algebra and taking the limit $\rho \to \infty$ the lemma follows easily. ■

Proposition 74 *[Full Conditional Posterior of Coefficients* $\binom{\mathbf{d}}{\mathbf{c}}$*. Version 2.]*
Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ *be observed, with* $\sigma^2 > 0$ *known, and set* $\mathbf{d} := (d_1\, d_2\, \cdots\, d_l)^\intercal$, $\mathbf{c} := (c_1\, c_2\, \cdots\, c_k)^\intercal$. *Consider the model*

$$\begin{aligned} y_i &= \sum_{j=1}^{l} d_j \phi_j(\mathbf{x}_i) + \sum_{j=1}^{k} c_j \psi_j(\mathbf{x}_i) + \epsilon_i, \\ \epsilon_i &\overset{iid}{\sim} N_1(0, \sigma^2), \end{aligned}$$

where $\{\phi\}_{i=1}^l$ *and* $\{\psi\}_{i=1}^k$ *are known functions. For* $b > 0, \tau > 0$ *and* $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ *positive definite matrix, consider the priors*

$$\begin{aligned} d_i &\overset{iid}{\sim} 1 \\ \mathbf{c} &\sim N_l(\mathbf{0}, b\,Q) \\ \mathbf{d} &\perp \mathbf{c} \\ \binom{\mathbf{d}}{\mathbf{c}} &\perp (\epsilon_1 \cdots \epsilon_n)^\intercal. \end{aligned}$$

Then the posterior of $\binom{\mathbf{d}}{\mathbf{c}}$ *is* $N_{l+k}(\mu_{\mathbf{dc}}, b\boldsymbol{\Sigma}_{\mathbf{dc}})$ *where*

$$\mu_{\mathbf{dc}} = \begin{pmatrix} (S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1} \\ QR^\intercal M^{-1}(I - S(S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1}) \end{pmatrix}\mathbf{y} \tag{B.15}$$

$$\boldsymbol{\Sigma}_{\mathbf{dc}} = \begin{pmatrix} (S^\intercal M^{-1}S)^{-1} & -(S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1}RQ \\ -QR^\intercal M^{-1}S(S^\intercal M^{-1}S)^{-1} & Q - QR^\intercal\{M^{-1} - M^{-1}S(S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1}\}RQ \end{pmatrix} \tag{B.16}$$

and $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ *with entry* $S_{i,j} = \phi_j(\mathbf{x}_i)$, $R \in \mathcal{M}_{n\times k}(\mathbb{R})$ *with entries* $R_{i,j} = \psi_j(\mathbf{x}_i)$, $M = RQR^\intercal + n\lambda I_n$, $n\lambda = \frac{\sigma^2}{b}$.

Proof.

First we compute the limits as $\rho \to \infty$ of the mean and covariance matrix (B.12) and (B.13).

Introducing the limits in the column vector and using Lemma 73 we have

$$\begin{aligned}
\lim_{\rho\to\infty}\mu_{\mathbf{dc}_\rho} &= \lim_{\rho\to\infty}\begin{pmatrix}\rho S^\intercal\left(\rho SS^\intercal+M\right)^{-1}\\ QR^\intercal\left(\rho SS^\intercal+M\right)^{-1}\end{pmatrix}\mathbf{y}\\
&= \begin{pmatrix}\lim_{\rho\to\infty}\rho S^\intercal\left(\rho SS^\intercal+M\right)^{-1}\\ \lim_{\rho\to\infty}QR^\intercal\left(\rho SS^\intercal+M\right)^{-1}\end{pmatrix}\mathbf{y}\\
&= \begin{pmatrix}\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\\ QR^\intercal M^{-1}\left(I_n-S\left(S^\intercal M^{-1}S\right)^{-1}S^\intercal M^{-1}\right)\end{pmatrix}\mathbf{y}\\
&= \mu_{\mathbf{dc}},
\end{aligned} \tag{B.17}$$

while for the covariance, matrix we proceed in a similar way introducing the limits inside the matrix and using Lemma 73:

$$\begin{aligned}
\lim_{\rho\to\infty}\mathbf{\Sigma}_{\mathbf{dc}_\rho} &= \lim_{\rho\to\infty} b\begin{pmatrix}\left(\rho I_k-\rho S^\intercal\left(\rho SS^\intercal+M\right)^{-1}\rho S\right) & \left(-\rho S^\intercal\left(\rho SS^\intercal+M\right)^{-1}RQ\right)\\ \left(-QR^\intercal\left(\rho SS^\intercal+M\right)^{-1}\rho S\right) & \left(Q-QR^\intercal\left(\rho SS^\intercal+M\right)^{-1}RQ\right)\end{pmatrix}\\
&= b\begin{pmatrix}(S^\intercal M^{-1}S)^{-1} & -(S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1}RQ\\ -QR^\intercal M^{-1}S(S^\intercal M^{-1}S)^{-1} & Q-QR^\intercal\{M^{-1}-M^{-1}S(S^\intercal M^{-1}S)^{-1}S^\intercal M^{-1}\}RQ\end{pmatrix}\\
&= b\mathbf{\Sigma}_{\mathbf{dc}}.
\end{aligned} \tag{B.18}$$

On the other side, the characteristic function of a multivariate normal distribution with mean $\mu_{\mathbf{dc}_\rho}$ and covariance $\mathbf{\Sigma}_{\mathbf{dc}_\rho}$ is $\phi_\rho(\mathbf{t}) = \exp\left(i\mathbf{t}^\intercal\mu_{\mathbf{dc}_\rho}+\frac{1}{2}b\mathbf{t}^\intercal\mathbf{\Sigma}_{\mathbf{dc}_\rho}\mathbf{t}\right)$ and $\mathbf{t}\in\mathbb{R}^{k+l}$. For fixed $\mathbf{t}$, taking the limits and using (*B*.17) and (*B*.18) we have that

$$\begin{aligned}
\lim_{\rho\to\infty}\phi_\rho(\mathbf{t}) &= \lim_{\rho\to\infty}\exp\left(i\mathbf{t}^\intercal\mu_{\mathbf{dc}_\rho}+\frac{1}{2}b\mathbf{t}^\intercal\mathbf{\Sigma}_{\mathbf{dc}_\rho}\mathbf{t}\right)\\
&= \exp\left(i\mathbf{t}^\intercal\left(\lim_{\rho\to\infty}\mu_{\mathbf{dc}_\rho}\right)+\frac{1}{2}b\mathbf{t}^\intercal\left(\lim_{\rho\to\infty}\mathbf{\Sigma}_{\mathbf{dc}_\rho}\right)\mathbf{t}\right)\\
&= \exp\left(i\mathbf{t}^\intercal\mu_{\mathbf{dc}}+\frac{1}{2}\mathbf{t}^\intercal\left(b\mathbf{\Sigma}_{\mathbf{dc}}\right)\mathbf{t}\right),
\end{aligned} \tag{B.19}$$

where expression (*B*.19) is the characteristic function of a random vector with distribution $N_{k+l}\left(\mu_{\mathbf{dc}}, b\mathbf{\Sigma}_{\mathbf{dc}}\right)$. By the Lévy's continuity Theorem (Athreya and Lahiri (2006)), (Fristedt and Gray (2013)), a sequence of random vectors $\left\{\binom{\mathbf{d}}{\mathbf{c}}_\rho\right\}_{\rho=1}^{\infty}$ such $\binom{\mathbf{d}}{\mathbf{c}}_\rho \sim N_{k+l}\left(\mu_{\mathbf{dc}_\rho}, b\mathbf{\Sigma}_{\mathbf{dc}_\rho}\right)$ converge in distribution as $\rho\to\infty$ to $\binom{\mathbf{d}}{\mathbf{c}} \sim N_{k+l}\left(\mu_{\mathbf{dc}}, b\mathbf{\Sigma}_{\mathbf{dc}}\right)$. We have shown that if in the

prior of $d_i \overset{iid}{\sim} N_1(0,\tau^2)$ in Lemma 72 we let $\frac{\tau^2}{b} = \rho \to \infty$, the posterior distribution (or the cumulative distribution function) of $\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)_\rho$ converges to $N_{k+l}\left(\mu_{\mathbf{dc}}, b\mathbf{\Sigma_{dc}}\right)$. Therefore, if we take $d_i \overset{iid}{\sim} 1$ as prior for d_i the posterior $\left[\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)\middle|\, \mathbf{y}\right]$ is $N_{k+l}\left(\mu_{\mathbf{dc}}, b\mathbf{\Sigma_{dc}}\right)$. ■

Proposition 75 (Equivalence Bayesian Models Auxiliary)

In the context of Lemma 74 we have $\left(\begin{smallmatrix}\mathbf{d}\\ \mathbf{c}\end{smallmatrix}\right)|\mathbf{y} \sim N_{l+k}\left(\mu_{\mathbf{dc}}, \mathbf{\Sigma_{dc}}\right)$, $\mu_{\mathbf{dc}}$ *and* $\mathbf{\Sigma_{dc}}$ *described by* (B.15) *and* (B.16). *Define* $\eta : \mathbb{X} \to \mathbb{R}$ *as*

$$\eta(x) := \sum_{j=1}^{l} d_j \phi_j(x) + \sum_{j=1}^{k} c_j \psi_j(x)$$

then $\eta|\mathbf{y}$ *is a Gaussian process with mean and covariance*

$$\mathbb{E}\left[\eta(x)|\mathbf{y}\right] = \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal} \mu_{\mathbf{dc}} \tag{B.20}$$

$$\begin{aligned} b^{-1}Cov\left(\eta(x),\eta(y)|\mathbf{y}\right) =& \Psi(x)^{\intercal}Q\Psi(y) + \Phi(x)^{\intercal}\left(S^{\intercal}M^{-1}S^{\intercal}\right)^{-1}\Phi(x) \\ &- \left[\Phi(x)^{\intercal}\tilde{\mathbf{d}}(y) + \Phi(y)^{\intercal}\tilde{\mathbf{d}}(x)\right] - \Psi(x)^{\intercal}\tilde{\mathbf{c}}(y) \end{aligned} \tag{B.21}$$

where

$$\begin{aligned} \Phi(x) &= (\phi_1(x)\cdots\phi_l(x))^{\intercal}, \\ \Psi(x) &= (\psi_1(x)\cdots\psi_k(x))^{\intercal}, \\ \tilde{\mathbf{d}}(x) &= \left(S^{\intercal}M^{-1}S\right)^{-1}S^{\intercal}M^{-1}RQ\Psi(x), \\ \tilde{\mathbf{c}}(x) &= QR^{\intercal}\left(M^{-1} - M^{-1}S\left(S^{\intercal}M^{-1}S\right)^{-1}S^{\intercal}M^{-1}\right)RQ\Psi(x). \end{aligned}$$

In particular

$$b^{-1}Var\left(\eta(x)|\mathbf{y}\right) = \Psi(x)^{\intercal}Q\Psi(x) + \Phi(x)^{\intercal}\left(S^{\intercal}M^{-1}S^{\intercal}\right)^{-1}\Phi(x) - 2\Phi(x)^{\intercal}\tilde{\mathbf{d}}(x) - \Psi(x)^{\intercal}\tilde{\mathbf{c}}(x). \tag{B.22}$$

Proof.

Consider

$$\begin{pmatrix}\mathbf{d}^{\star}\\ \mathbf{c}^{\star}\end{pmatrix} = \begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} - \mu_{\mathbf{dc}}$$

$$\eta^{\star}(x) = \sum_{j=1}^{l} d_j^{\star}\phi_j(x) + \sum_{j=1}^{k} c_j^{\star}\psi_j(x).$$

By Lemma 71, $\eta^{\star}$ is a Gaussian process with mean 0 and covariance $Cov(\eta^{\star}(x), \eta^{\star}(y)) = \binom{\Phi(x)}{\Psi(x)}^{\intercal}\mathbf{\Sigma_{dc}}\binom{\Phi(y)}{\Psi(y)}$. Therefore, we have

$$\begin{aligned}
b^{-1}Cov(\eta(x),\eta(y)) =& \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\begin{pmatrix}(S^{\intercal}M^{-1}S)^{-1} & -(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ\\ -QR^{\intercal}M^{-1}S(S^{\intercal}M^{-1}S)^{-1} & Q-QR^{\intercal}\{M^{-1}-M^{-1}S(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}\}RQ\end{pmatrix}\begin{pmatrix}\Phi(y)\\ \Psi(y)\end{pmatrix}\\
=&\Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}\Phi(y) - \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ\Psi(y)\\
&-\Psi(x)^{\intercal}QR^{\intercal}M^{-1}S(S^{\intercal}M^{-1}S)^{-1}\Phi(y)\\
&+\Psi(x)^{\intercal}\left[Q - QR^{\intercal}\{M^{-1}-M^{-1}S(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}\}RQ\right]\Psi(y)\\
=&\Psi(x)^{\intercal}Q\Psi(y) + \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}\Phi(y) - \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ\Psi(y)\\
&-\left[\Phi(y)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}S^{\intercal}M^{-1}RQ\Psi(x)\right]^{\intercal}\\
&-\Psi(x)QR^{\intercal}\left(M^{-1}-M^{-1}S\left(S^{\intercal}M^{-1}S\right)^{-1}S^{\intercal}M^{-1}\right)RQ\Psi(y)\\
=&\Psi(x)^{\intercal}Q\Psi(y) + \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}\Phi(y) - \phi(x)^{\intercal}\tilde{\mathbf{d}}(y) - \left[\phi(y)^{\intercal}\tilde{\mathbf{d}}(x)\right]^{\intercal} - \Psi(x)^{\intercal}\tilde{\mathbf{c}}(y)\\
=&\Psi(x)^{\intercal}Q\Psi(y) + \Phi(x)^{\intercal}(S^{\intercal}M^{-1}S)^{-1}\Phi(y) - \left[\phi(x)^{\intercal}\tilde{\mathbf{d}}(y) + \phi(y)^{\intercal}\tilde{\mathbf{d}}(x)\right] - \Psi(x)^{\intercal}\tilde{\mathbf{c}}(y).
\end{aligned}$$

Since $\eta(x) = \eta^{\star}(x) + \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}$ we have

$$\mathbb{E}[\eta(x)|\mathbf{y}] = \mathbb{E}\left[\eta^{\star}(x) + \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}\middle|\mathbf{y}\right] = \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}$$

and

$$Cov(\eta(x),\eta(y)|\mathbf{y}) = Cov\left(\eta^{\star}(x) + \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}, \eta^{\star}(y) + \begin{pmatrix}\Phi(x)\\ \Psi(x)\end{pmatrix}^{\intercal}\mu_{\mathbf{dc}}\middle|\mathbf{y}\right) = Cov(\eta^{\star}(x),\eta^{\star}(y)|\mathbf{y}).$$

The expression for $Var(\eta|\mathbf{y})$ follows directly ■

B.2 Vector Valued Functions in Hilbert Spaces

Proposition 76

In the context of Proposition 25

$$\mathbf{\Gamma\Gamma}^{+}\mathbf{K}_{zx} = \mathbf{K}_{zx}.$$

Proof.

Lets define

$$\xi(\mathbf{x}) = \begin{pmatrix} \mathcal{K}(\mathbf{z}_1, \mathbf{x}) \\ \vdots \\ \mathcal{K}(\mathbf{z}_k, \mathbf{x}) \end{pmatrix} \in \mathcal{M}_{kd_2 \times d_2}(\mathbb{R}).$$

First I want to prove that for any $\mathbf{x} \in \mathbb{R}^{d_1}$, every column of $\xi(\mathbf{x})$ is in $Im(\mathbf{\Gamma})$. Let be $\mathbb{A} = \{\mathbf{a} : \mathbf{a}^\intercal(\mathbf{\Gamma})\mathbf{a} = 0\}$. Let be $\mathbf{x} \in \mathbb{R}^{kd_2}$ and $\mathbf{a} \in \mathbb{A}$. Observe that $0 = \mathbf{a}^\intercal(\mathbf{\Gamma})\mathbf{a} = J(\xi(\mathbf{x})^\intercal\mathbf{a})$ implies $\xi(\mathbf{x})^\intercal\mathbf{a} = 0$ because $\|\cdot\|^2_{\mathcal{HK}}$ is a norm in $span\{\mathcal{K}(\mathbf{z}_i, \cdot)\}_{i=1}^k$. Hence, every column of $\xi(\mathbf{x})$ is in an orthogonal-$\langle\cdot,\cdot\rangle_{\mathcal{H}_\mathcal{K}}$ space to $\mathbb{A}$; $\xi(\mathbf{x}) \in \mathbb{A}^\intercal$.

On another side, from definitions we have that $ker(\mathbf{\Gamma}) \subset \mathbb{A}$, then $Im(\mathbf{\Gamma})) = (ker(\mathbf{\Gamma}))^\perp \supseteq \mathbb{A}^\perp$. Thus we can conclude that every column of $\xi(\mathbf{x})$ is in $Im(\mathbf{\Gamma})$.

From before and in particular, every column of $\mathbf{K}_{zx}$ is in the image of $\mathbf{\Gamma}$. We also know by definition of the Moore-Penrose inverse that $\mathbf{\Gamma}\mathbf{\Gamma}^+\mathbf{\Gamma} = \mathbf{\Gamma}$, hence $\mathbf{\Gamma}\mathbf{\Gamma}^+$ projects every vector in the column space of $\mathbf{\Gamma}$, the image of $\mathbf{\Gamma}$, then $\mathbf{\Gamma}\mathbf{\Gamma}^+$ leaves fixed any vector in the image of $\mathbf{\Gamma}$ such as every column of $\mathbf{K}_{zx}$, then $\mathbf{\Gamma}\mathbf{\Gamma}^+\mathbf{K}_{zx} = \mathbf{K}_{zx}$. ■

Proposition 77

Let $\mathcal{H}$ be a RKHS of functions with domain in $\mathbb{R}^{d_1}$ and rank in $\mathbb{R}^{d_2}$. Let $\langle\cdot,\cdot\rangle$ be the inner product in $\mathcal{H}$ and $\|\cdot\|$ its induced norm. Let $\{\mathbf{z}_i\}_{i=1}^k, \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^{d_1}$, $\Sigma \in \mathcal{M}_{d_2 \times d_2}(\mathbb{R})$ a positive definite matrix. Let $\eta \in \mathcal{H}$ be of the form

$$\eta(\mathbf{x}) = \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i, \mathbf{x})\mathbf{a}_i. \tag{B.23}$$

Then, the function

$$\mathfrak{L}(\mathcal{A}) := \sum_{i=1}^n (\mathbf{y}_i - \eta(\mathbf{x}_i))^\intercal \Sigma^{-1} (\mathbf{y}_i - \eta(\mathbf{x}_i)) + n\lambda\|\eta\|^2_{\mathcal{H}}$$

can be written as

$$\mathfrak{L}(\mathcal{A}) = (\mathbf{Y} - \mathbf{K}_{xz}\mathcal{A})^\intercal \Psi_{\Sigma,n} (\mathbf{Y} - \mathbf{K}_{xz}\mathcal{A}) + \frac{n\lambda}{2}\mathcal{A}^\intercal\mathbf{\Gamma}\mathcal{A}, \tag{B.24}$$

where $\mathbf{Y} = \mathrm{vec}(\mathbf{y}_1 \cdots \mathbf{y}_n)$, $\mathcal{A} = \mathrm{vec}(\mathbf{a}_1 \cdots \mathbf{a}_k)$, $\mathbf{K}_{xz} \in \mathcal{M}_{nd_2 \times kd_2}(\mathbb{R})$ is a block matrix with i, jth block $\mathcal{K}(\mathbf{x}_i, \mathbf{z}_j)$, $\mathbf{K}_{zz} \in \mathcal{M}_{kd_2 \times xd_2}(\mathbb{R})$ is another block matrix with i, jth block $\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)$,

$\mathbf{K}_{zx} = \mathbf{K}_{xz}^{\intercal}$, $\Psi_{\Sigma,n} = I_n \otimes \Sigma^{-1}$, $\mathbf{\Gamma} = \mathbf{K}_{zz}\Psi_{\Sigma,n} + \Psi_{\Sigma,n}\mathbf{K}_{zz}$. *Furthermore, the gradient and Hessian matrix are*

$$\nabla \mathfrak{L}(\mathcal{A})^{\intercal} = 2\left\{\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + \frac{n\lambda}{2}\mathbf{\Gamma}\right\}\mathcal{A} - 2\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{Y} \text{ and}$$
$$\mathbf{H}\left[\mathfrak{L}(\mathcal{A})\right] = 2\left\{\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + \frac{n\lambda}{2}\mathbf{\Gamma}\right\}$$

respectively, the critical points of $\mathfrak{L}$ *in* $\mathbb{R}^{kd_2}$ *satisfy the linear equations*

$$\left\{\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + \frac{n\lambda}{2}\mathbf{\Gamma}\right\}\mathcal{A} = \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{Y}, \tag{B.25}$$

any critical point is not a local minimum; and any two critical points $\mathcal{A}$ *and* $\mathcal{X}$ *have the property that* $\mathfrak{L}(\mathcal{A}) = \mathfrak{L}(\mathcal{X})$.

Proof.

Let us first prove that $\mathfrak{L}$ can be written as (B.24). Lets compute $\|\eta\|_{\mathcal{H}}^2$, with η as expression (B.23):

$$\begin{aligned}
\|\eta\|_{\mathcal{H}}^2 &= \left\langle \sum_{i=1}^{k} \mathcal{K}(\mathbf{z}_i, \cdot)\mathbf{a}_i, \sum_{i=1}^{k} \mathcal{K}(\mathbf{z}_i, \cdot)\mathbf{a}_j \right\rangle_{\mathcal{H}_{\mathcal{K}}} \\
&= \sum_{i,j=1}^{k} \langle \mathcal{K}(\mathbf{z}_i, \cdot)\mathbf{a}_i, \mathcal{K}(\mathbf{z}_j, \cdot)\mathbf{a}_j \rangle_{\mathcal{H}} \\
&= \frac{1}{2}\left\{ \sum_{i,j=1}^{k} \langle \mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\mathbf{a}_i, \mathbf{a}_j \rangle_{\mathbb{Y}} + \sum_{i,j=1}^{k} \langle \mathbf{a}_i, \mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\mathbf{a}_j \rangle_{\mathbb{Y}} \right\} \qquad \text{(B.26)} \\
&= \frac{1}{2}\left\{ \sum_{i,j=1}^{k} \mathbf{a}_i^{\intercal}\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\Sigma^{-1}\mathbf{a}_j + \sum_{i,j=1}^{k} \mathbf{a}_i^{\intercal}\Sigma^{-1}\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\mathbf{a}_j \right\} \\
&= \frac{1}{2}\sum_{i,j=1}^{k} \mathbf{a}_i^{\intercal}\left\{\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\Sigma^{-1} + \Sigma^{-1}\mathcal{K}(\mathbf{z}_i, \mathbf{z}_j)\right\}\mathbf{a}_j \\
&= \frac{1}{2}\mathcal{A}\left[\mathbf{K}_{zz}\Psi_{\Sigma,n} + \Psi_{\Sigma,n}\mathbf{K}_{zz}\right]\mathcal{A}, \\
&= \frac{1}{2}\mathcal{A}\mathbf{\Gamma}\mathcal{A}, \qquad \text{(B.27)}
\end{aligned}$$

where (B.26) is obtained using the reproducing property and symmetry of the inner products. Furthermore, it is direct to see that

$$\sum_{i=1}^{n} \left(\mathbf{y}_i - \eta(\mathbf{x}_i)\right)^{\intercal} \Sigma^{-1} \left(\mathbf{y}_i - \eta(\mathbf{x}_i)\right) = \left(\mathbf{Y} - \mathbf{K}_{xz}\mathcal{A}\right)^{\intercal} \Psi_{\Sigma,n} \left(\mathbf{Y} - \mathbf{K}_{xz}\mathcal{A}\right), \tag{B.28}$$

hence, by (*B*.27) and (*B*.28), we have proven (*B*.24). Expression (*B*.25) is obtained directly by equating the corresponding quantities.

Lets obtain the gradient $\nabla\mathfrak{L}$ computing the derivative directly. Let $\mathbf{Y}^{\star} = (\Psi_{\Sigma,n})^{1/2}\mathbf{Y} = \Psi_{\Sigma^{1/2},n}\mathbf{Y}$ and $\mathbf{K}_{xz}^{\star} = (\Psi_{\Sigma,n})^{1/2}\mathbf{K}_{xz} = \Psi_{\Sigma^{1/2},n}\mathbf{K}_{xz}$ (Definition 33 for the square root of a matrix). Expanding (*B*.24) we have

$$\mathfrak{L}(\mathcal{A}) = (\mathbf{Y}^{\star})^{\intercal}\mathbf{Y}^{\star} - \mathcal{A}^{\intercal}(\mathbf{K}_{xz}^{\star})^{\intercal}\mathbf{Y}^{\star} - (\mathbf{Y}^{\star})^{\intercal}\mathbf{K}_{xz}^{\star}\mathcal{A} + \mathcal{A}^{\intercal}(\mathbf{K}_{xz}^{\star})^{\intercal}\mathbf{K}_{xz}^{\star}\mathcal{A} + \frac{n\lambda}{2}\mathcal{A}^{\intercal}\mathbf{\Gamma}\mathcal{A},$$

taking derivatives we have

$$\begin{aligned}
\nabla\mathfrak{L}(\mathcal{A}) &= -2(\mathbf{Y}^{\star})^{\intercal}\mathbf{K}_{xz}^{\star} + 2\mathcal{A}^{\intercal}(\mathbf{K}_{xz}^{\star})^{\intercal}\mathbf{K}_{xz}^{\star} + 2\frac{n\lambda}{2}\mathcal{A}^{\intercal}\mathbf{\Gamma} \\
&= -2\mathbf{Y}^{\intercal}\Psi_{\Sigma,n}\mathbf{K}_{xz} + 2\mathcal{A}^{\intercal}\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + 2\frac{n\lambda}{2}\mathcal{A}^{\intercal}\mathbf{\Gamma} \\
&= 2\left[\left\{\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + \frac{n\lambda}{2}\mathbf{\Gamma}\right\}\mathcal{A} - \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{Y}\right]^{\intercal},
\end{aligned}$$

and the Jacobian matrix is directly obtained as

$$\mathbf{H}\left[\mathfrak{L}(\mathcal{A})\right] = 2\left\{\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + \frac{n\lambda}{2}\mathbf{\Gamma}\right\}.$$

Expression (*B*.25) is obtained directly by equation $\nabla\mathfrak{L}(\mathcal{A}) = \mathbf{0}$.

Now, we prove that $\mathfrak{L}(\mathcal{X}) = \mathfrak{L}(\mathcal{A})$ for any two critical points $\mathcal{X}, \mathcal{A} \in \mathbb{R}^{N}$. Observe that using the Taylor approximation Theorem, centering the approximation in $\mathcal{A}$ and that the third partial derivatives are zero, $\mathfrak{L}$ can be written as

$$\mathfrak{L}(\mathcal{X}) = \mathfrak{L}(\mathcal{A}) + \nabla\mathfrak{L}(\mathcal{A})^{\intercal}(\mathcal{X} - \mathcal{A}) + (\mathcal{X} - \mathcal{A})^{\intercal}\mathbf{H}\left[\mathfrak{L}(\mathcal{A})\right](\mathcal{X} - \mathcal{A}).$$

Furthermore, by Proposition 78 the Hessian matrix $\mathbf{H}\left[\mathfrak{L}(\mathcal{A})\right]$ is semipositive definite in any critical point $\mathcal{A}$, then

$$\begin{aligned}
\mathfrak{L}(\mathcal{X}) &= \mathfrak{L}(\mathcal{A}) + (\mathcal{X} - \mathcal{A})^{\intercal}\mathbf{H}\left[\mathfrak{L}(\mathcal{A})\right](\mathcal{X} - \mathcal{A}) \geq \mathfrak{L}(\mathcal{A}) \\
\mathfrak{L}(\mathcal{A}) &= \mathfrak{L}(\mathcal{X}) + (\mathcal{A} - \mathcal{X})^{\intercal}\mathbf{H}\left[\mathfrak{L}(\mathcal{A})\right](\mathcal{A} - \mathcal{X}) \geq \mathfrak{L}(\mathcal{X}),
\end{aligned}$$

then $\mathfrak{L}(\mathcal{A}) = \mathfrak{L}(\mathcal{X})$ for any two critical points of $\mathfrak{L}$. Furthermore, we also have by the previous equations that any critical point is not a local minimum.

We have proven that any critical point $\mathcal{X}$ is a global maximum even when it may be another $\mathcal{A}$ such $\mathfrak{L}(\mathcal{X}) = \mathfrak{L}(\mathcal{A})$. ■

Proposition 78

In the context of Proposition 77, the $\mathcal{M}_{kd_2\times kd_2}(\mathbb{R})$ *matrix*

$$\{\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + n\lambda\mathbf{\Gamma}\}$$

is semipositive definite in the usual sense of matrices.

Proof.

First we prove $\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}$ is semipositive definite matrix. Let $\mathcal{A} \in \mathbb{R}^{kd_2}$ and observe that

$$\mathcal{A}^\intercal\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\mathcal{A} = [\mathbf{K}_{xz}\mathcal{A}]^\intercal\,\Psi_{\Sigma,n}\,[\mathbf{K}_{xz}\mathcal{A}] \geq 0,$$

where the last inequality is because $\Psi_{\Sigma,n}$ is positive definite. Hence $\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}$ is semipositive definite matrix.

Second, we proof that $\mathbf{\Gamma} = \mathbf{K}_{zz}\Psi_{\Sigma,n} + \Psi_{\Sigma,n}\mathbf{K}_{zz}$ is semipositive definite matrix. Let $\mathcal{A} = \text{vec}\,(\mathbf{a}_1 \cdots \mathbf{a}_k)$ and $\eta = \sum_{i=1}^n \mathcal{K}(\mathbf{z}_i, \cdot)\mathbf{a}_i$. Observe that

$$\begin{aligned}
0 \leq \|\eta\|^2 &= \left\langle \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i,\cdot)\mathbf{a}_i, \sum_{i=1}^k \mathcal{K}(\mathbf{z}_i,\cdot)\mathbf{a}_j \right\rangle_{\mathcal{H}_\mathcal{K}} \\
&= \sum_{i,j=1}^k \langle \mathcal{K}(\mathbf{z}_i,\cdot)\mathbf{a}_i, \mathcal{K}(\mathbf{z}_j,\cdot)\mathbf{a}_j\rangle_{\mathcal{H}_\mathcal{K}} \\
&= \frac{1}{2}\left\{ \sum_{i,j=1}^k \langle \mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\mathbf{a}_i, \mathbf{a}_j\rangle_{\mathbb{Y}} + \sum_{i,j=1}^k \langle \mathbf{a}_i, \mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\mathbf{a}_j\rangle_{\mathbb{Y}} \right\} \qquad \text{(B.29)} \\
&= \frac{1}{2}\left\{ \sum_{i,j=1}^k \mathbf{a}_i^\intercal \mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\Sigma^{-1}\mathbf{a}_j + \sum_{i,j=1}^k \mathbf{a}_i^\intercal \Sigma^{-1}\mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\mathbf{a}_j \right\} \\
&= \frac{1}{2}\sum_{i,j=1}^k \mathbf{a}_i^\intercal \left\{\mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\Sigma^{-1} + \Sigma^{-1}\mathcal{K}(\mathbf{z}_i,\mathbf{z}_j)\right\}\mathbf{a}_j \\
&= \frac{1}{2}\mathcal{A}\,[\mathbf{K}_{zz}\Psi_{\Sigma,n} + \Psi_{\Sigma,n}\mathbf{K}_{zz}]\,\mathcal{A},
\end{aligned}$$

where (*B.29*) is obtained using the reproducing property and symmetry of the inner products. Therefore, $\mathbf{\Gamma}$ is a semipositive definite matrix. Hence, $n\lambda\mathbf{\Gamma}$ is semipositive definite because $n\lambda > 0$

We have shown that $\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}$ and $n\lambda\mathbf{\Gamma}$ are semipositive definite matrices. The sum of two semipositive definite matrices is semipositive definite. We have proven what we wanted. ■

Proposition 79

In the context of Proposition 18 and Proposition 77, $\mu_{\mathbf{Y}}$ satisfies equation (B.25).

Proof.

Recall that $\Psi_{\Sigma,n} = \left(I_n \otimes \Sigma^{-1}\right)$, $\Psi_{\Sigma^{-1},n} = \left(I_n \otimes \Sigma\right)$ and let $\mathbf{M} = \mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma,n}$. Lets prove now that μ_Y satisfies (B.25) by plugging $\mu_{\mathbf{Y}} = \mathbf{\Gamma}^{+}\mathbf{K}_{xz}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}^{-1}\mathbf{Y}$ into left part of equation (B.25). We obtain:

$$\begin{aligned}
\left\{\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz} + \frac{n\lambda}{2}\mathbf{\Gamma}\right\}\mu_{\mathbf{Y}} &= \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx}\mathbf{M}^{-1}\mathbf{Y} + \frac{n\lambda}{2}\left(\mathbf{\Gamma}\mathbf{\Gamma}^{+}\mathbf{K}_{zx}\right)M^{-1}\mathbf{Y} \\
&= \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx}\mathbf{M}^{-1}\mathbf{Y} + \frac{n\lambda}{2}\left(\mathbf{K}_{zx}\right)M^{-1}\mathbf{Y} \qquad \text{(B.30)} \\
&= \mathbf{K}_{zx}\left[\Psi_{\Sigma,n}\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}I_{nd_2}\right]\mathbf{M}^{-1}\mathbf{Y} \\
&= \mathbf{K}_{zx}\left[\Psi_{\Sigma,n}\left\{\mathbf{K}_{xz}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\}M^{-1}\right]\mathbf{Y} \\
&= \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{Y},
\end{aligned}$$

where (B.30) is by Proposition 76 that states $\mathbf{\Gamma}\mathbf{\Gamma}^{+}\mathbf{K}_{zx} = \mathbf{K}_{zx}$. ■

Proposition 80 *Using the notation from Proposition 25, lets define*

$$\mathbf{M} = \mathbf{K}_{xz}\left[\Psi_{\Sigma,k}\mathbf{K}_{zz} + \mathbf{K}_{zz}\Psi_{\Sigma,k}\right]^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}.$$

Then,

$$\mathbf{M}^{-1} = \frac{2}{n\lambda}\Psi_{\Sigma,n}\left[\Psi_{\Sigma^{-1},n} - \mathbf{K}_{xz}\left(\frac{n\lambda}{2}\mathbf{K}_{zz} + \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\mathbf{K}_{zx}\right]\Psi_{\Sigma,n}. \qquad \text{(B.31)}$$

Proof.

Define $\mathbf{N}$ as

$$\mathbf{N} = \frac{2}{n\lambda}\Psi_{\Sigma,n}\left[\Psi_{\Sigma^{-1},n} - \mathbf{K}_{xz}\left(\frac{n\lambda}{2}\mathbf{K}_{zz} + \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\mathbf{K}_{zx}\right]\Psi_{\Sigma,n}.$$

We prove now that $\mathbf{MN} = I_n$ as

$$\begin{aligned}
\mathbf{MN} =& \frac{2}{n\lambda}\left\{\mathbf{K}_{xz}\left[\Psi_{\Sigma,k}\mathbf{K}_{zz} + \mathbf{K}_{zz}\Psi_{\Sigma,k}\right]^{+}\mathbf{K}_{zx} + \frac{n\lambda}{2}\Psi_{\Sigma^{-1},n}\right\} \\
&\times\left\{\Psi_{\Sigma,n} - \Psi_{\Sigma,n}\mathbf{K}_{xz}\left(\frac{n\lambda}{2}\mathbf{K}_{zz} + \mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\right\}
\end{aligned}$$

$$
\begin{aligned}
=&2(n\lambda)^{-1}\left\{\mathbf{K}_{xz}\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}-\mathbf{K}_{xz}\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\left(\frac{n\lambda}{2}\mathbf{K}_{zz}+\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\right\}\\
&\times 2(n\lambda)^{-1}\left\{\frac{n\lambda}{2}I_n-\mathbf{K}_{xz}\left(\frac{n\lambda}{2}\mathbf{K}_{zz}+\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\mathbf{K}_{zx}\right\}\\
=&I_n+2(n\lambda)^{-1}\mathbf{K}_{xz}\left\{\boldsymbol{\Gamma}^{+}-\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\left(\frac{n\lambda}{2}\mathbf{K}_{zz}+\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\right\}\mathbf{K}_{zx}\Psi_{\Sigma,n}\\
&\times 2(n\lambda)^{-1}\mathbf{K}_{xz}\left\{-\frac{n\lambda}{2}\left(\frac{n\lambda}{2}\mathbf{K}_{zz}+\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\right\}\mathbf{K}_{zx}\Psi_{\Sigma,n}\\
=&I_n+2(n\lambda)^{-1}\mathbf{K}_{xz}\left\{\boldsymbol{\Gamma}^{+}-\left(\frac{n\lambda}{2}I_k+\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)\left(\frac{n\lambda}{2}\boldsymbol{\Gamma}+\left(\mathbf{K}_{zx}\right)\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\right\}\mathbf{K}_{zx}\Psi_{\Sigma,n}\\
=&I_n+2(n\lambda)^{-1}\mathbf{K}_{xz}\left\{\boldsymbol{\Gamma}^{+}-\left(\frac{n\lambda}{2}I_k+\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)\left(\frac{n\lambda}{2}\boldsymbol{\Gamma}+\left(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\right)\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\right\}\mathbf{K}_{zx}\Psi_{\Sigma,n}
\end{aligned}
\tag{B.32}
$$

$$
\begin{aligned}
=&I_n+2(n\lambda)^{-1}\mathbf{K}_{xz}\left\{\boldsymbol{\Gamma}^{+}-\left(\frac{n\lambda}{2}I_k+\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)\left(\frac{n\lambda}{2}I_k+\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}\right)^{+}\boldsymbol{\Gamma}^{+}\right\}\mathbf{K}_{zx}\Psi_{\Sigma,n}\\
=&I_n+2(n\lambda)^{-1}\mathbf{K}_{xz}\left\{\boldsymbol{\Gamma}^{+}-\boldsymbol{\Gamma}^{+}\right\}\mathbf{K}_{zx}\Psi_{\Sigma,n}
\end{aligned}
\tag{B.33}
$$

$$
=I_n,
$$

where expression (*B.32*) using that $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}=\mathbf{K}_{zx}$ (Lemma 76), and (*B.33*) is because $\frac{n\lambda}{2}I_k+\boldsymbol{\Gamma}^{+}\mathbf{K}_{zx}\Psi_{\Sigma,n}\mathbf{K}_{xz}$ is invertible. Using a similar procedure with the equality $\mathbf{K}_{xz}\boldsymbol{\Gamma}^{+}\boldsymbol{\Gamma}=\mathbf{K}_{xz}$, we can prove that $\mathbf{NM}=I_n$. Therefore $\mathbf{M}^{-1}=\mathbf{N}$. ■

Theorem 81 (Representer Theorem I) *Micchelli and Pontil (2005)*

Let $\mathbb{X}$ a non empty set, $\mathbb{Y}$ a real Hilbert space with inner product $<\cdot,\cdot>_{\mathbb{Y}}$. Let $\mathcal{H}$ a RKHS of functions $\eta:\mathbb{X}\to\mathbb{Y}$. Let $\{\mathbf{x}_i,\mathbf{y}_i\}_{i=1}^{n}$ an observed labeled sample with $\mathbf{x}_i\in\mathbb{X}$ and $\mathbf{y}_i\in\mathbb{Y}$. The minimization problem

$$
\arg\min_{\eta\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\|\eta(\mathbf{x}_i)-\mathbf{y}_i\|_{\mathbb{Y}}^{2}+\lambda\|\eta\|_{\mathcal{H}}^{2}
$$

has a unique solution $\eta=\sum_{i=1}^{n}\mathcal{K}(\mathbf{x}_i,\cdot)\mathbf{a}_i$, where the vector $\mathbf{a}_i\in\mathbb{Y}$ satisfy the n linear equations

$$
\sum_{j=1}^{n}\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)a_j+n\lambda\mathbf{a}_i=\mathbf{y}_i,\ \textit{for } 1\leq i\leq n.
$$

Theorem 82 (Representer Theorem II) *Minh et al. (2013)*

Let $\mathbb{X}$ a non empty set, $\mathbb{Y}$ a real Hilbert space with inner product $< \cdot, \cdot >_{\mathbb{Y}}$. Let $\mathcal{H}$ a RKHS of functions $\eta : \mathbb{X} \to \mathbb{Y}$. Let M: $\mathbb{Y}^{n+m} \to \mathbb{Y}^{n+m} \in \mathfrak{L}(\mathbb{Y}^{n+m})$ a symmetric positive operator. Let $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{n} \cup \{\mathbf{x}_i\}_{i=n+1}^{n+m}$ an observed sample with $\mathbf{x}_i \in \mathbb{X}$ and $\mathbf{y}_i \in \mathbb{Y}$. The minimization problem

$$\arg\min_{\eta \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \|\eta(\mathbf{x}_i) - \mathbf{y}_i\|_{\mathbb{Y}}^2 + \lambda_A \|\eta\|_{\mathcal{H}}^2 + \lambda_I \langle \eta, M\eta \rangle_{\mathbb{Y}}$$

has a unique solution $\eta = \sum_{i=1}^{n+m} \mathcal{K}(\mathbf{x}_i, \cdot)\mathbf{a}_i$, where the vector $\mathbf{a}_i \in \mathbb{Y}$ satisfy the linear equations

$$\sum_{j=1}^{n+m} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) a_j + m\lambda_I \sum_{j,k=1}^{n+m} M_{ik} \mathcal{K}(\mathbf{x}_k, \mathbf{x}_j)\mathbf{a}_j + n\lambda_A \mathbf{a}_i = \mathbf{y}_i, \text{ for } 1 \leq i \leq n$$

$$\lambda_I \sum_{j,k=1}^{n+m} M_{ik} \mathcal{K}(\mathbf{x}_k, \mathbf{x}_j)\mathbf{a}_j + \lambda_A \mathbf{a}_i = \mathbf{0}, \text{ for } n+1 \leq i \leq n+m.$$

APPENDIX C. FIGURES

In this appendix, we present more complete summary plots from the simulation studies described in the main body of the dissertation.

C.1 Real Valued Regression Functions

We show plots that summarize Chapter 3; estimation of multivariate real regression functions. Each plot has its own description and reference to the main body of the dissertation.

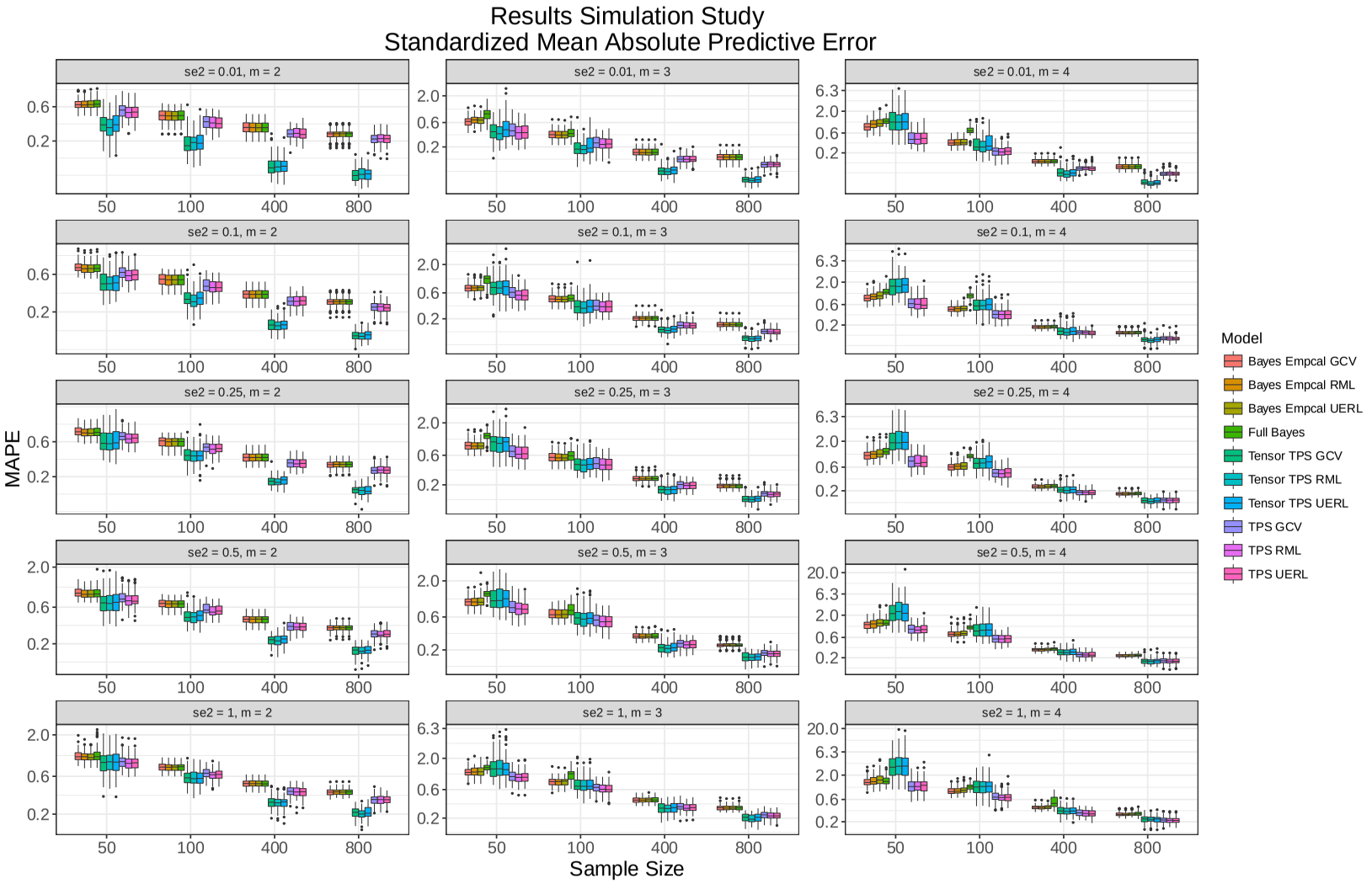

Figure C.1 Boxplots of the simulation results for the standardized mean absolute predictive error (3.19) (MAPE) for the multivariate regression problem predicting over the grid of resolution 0.05×0.05 inside the square $[-2.25, 2.25]^2$. Observe that the y-axis is in the log_{10} scale and each plot has different range of values. The rows indicates the true observation-error variance σ^2. The columns indicate the degree of derivative m for the penalty in (2.23). The models are described in Table 3.3.

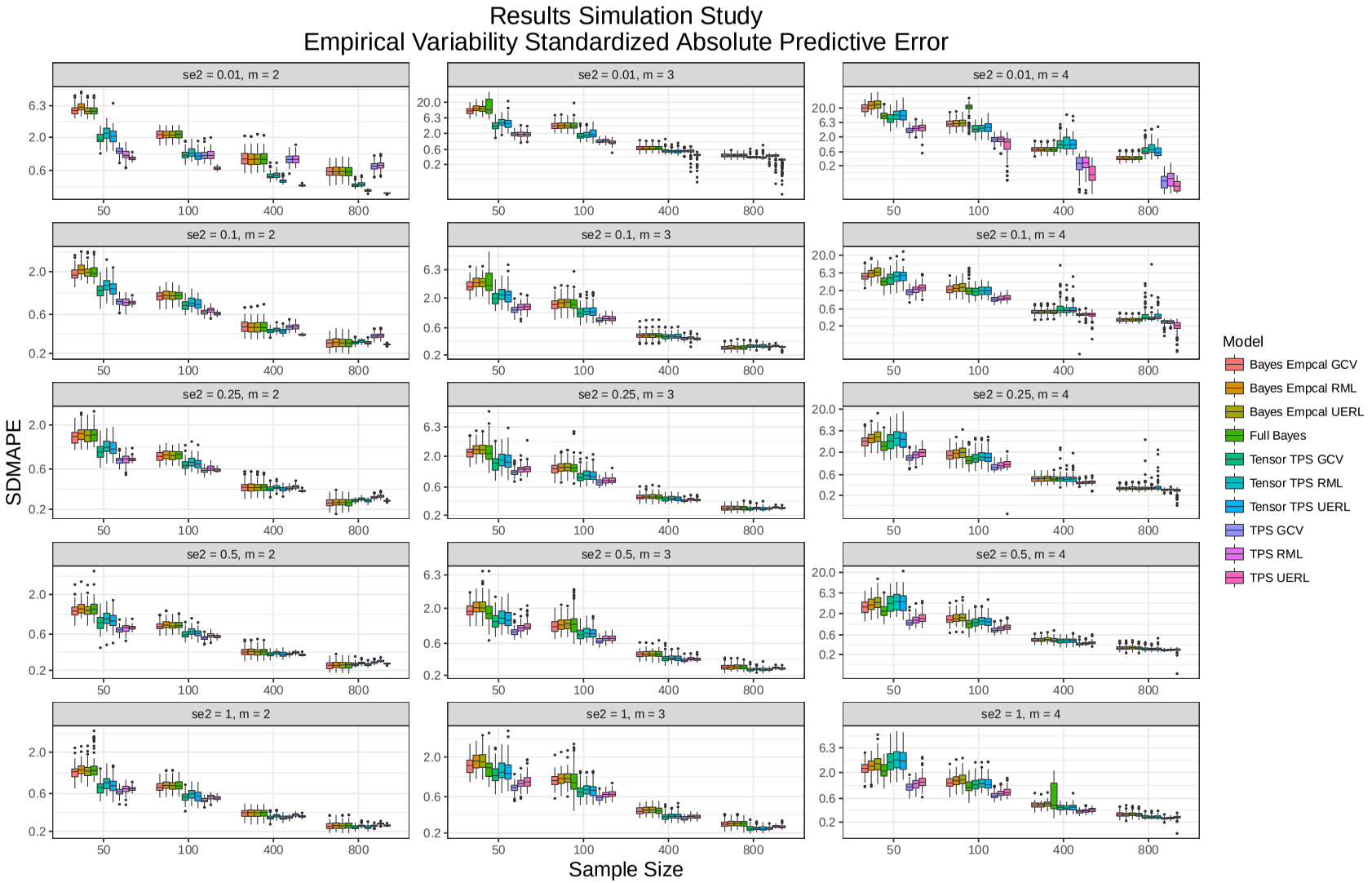

Figure C.2 Box plots to summarize the simulation results empirical variability absolute predictive error (3.20) for the multivariate regression problem.Each box plot is formed with the SDMAPE from 200 simulated data sets and the prediction of $\eta(\chi_i)$ is over the square $[-2.25, 2.25]^2$ in a grid 0.05×0.05. Observe that the y-axis is in the log_{10} scale and each plot has a different range of values on the vertical axis.

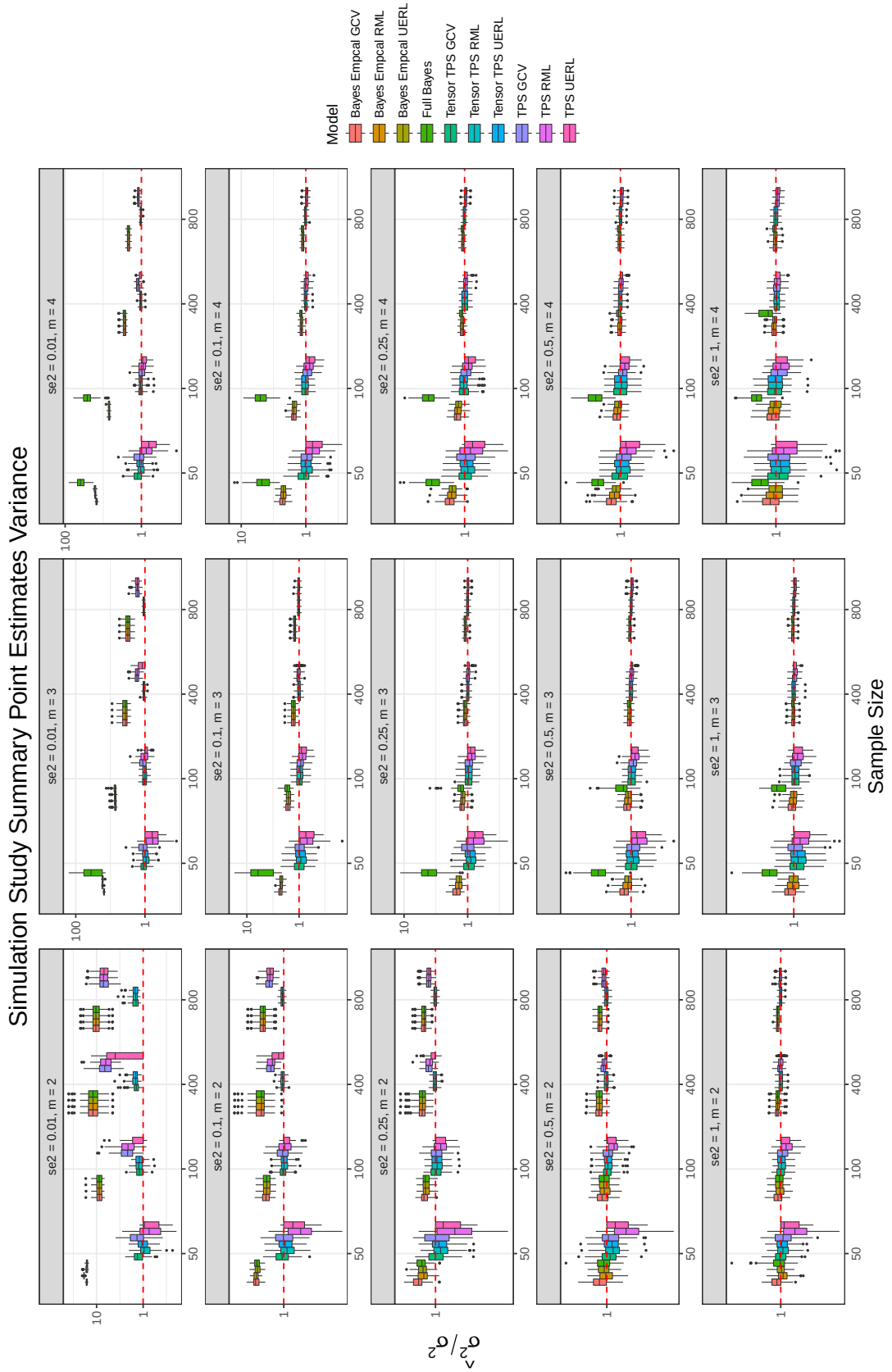

Figure C.3 Mean marginal posterior of the observation-error variance parameter σ^2. Each column indicate the true observation-error variance σ^2. The models are described in Table 3.3. Observe that in the $y-axis$ is plotted the ratio of the estimated variance and the true variance; it is desired to have a ratio of 1, furthermore, the y-axis is in the log_{10} scale and each plot has a different range of values.

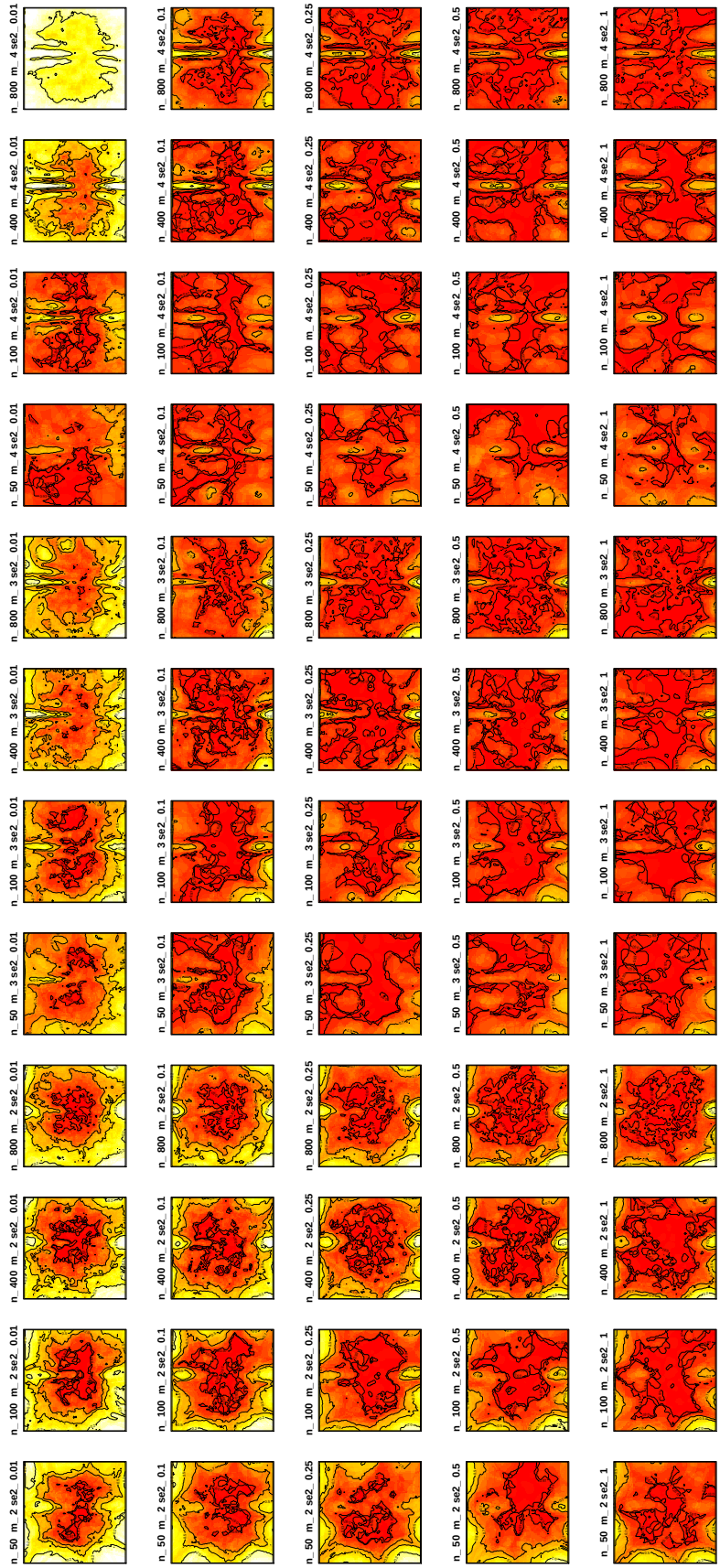

Figure C.4 Level curves of the empirical coverage for the pointwise 95% credible intervals using the Bayesian model with thin plate splines and $m = 3$. Smoothing parameter chosen with the restricted maximum likelihood method. Each value in the level plot is an estimate of ρ_i for the coverage of $\eta(\chi_i)$ in the grid $\{\chi_i\}_{i=1}^{N}$ using 200 different simulated data sets and computing the respective credible intervals. Each data set was simulated with $n = 100$, $\sigma^2 = 0.5$ and $\mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$. The target function is described by (3.17) and plotted in Figure 3.1. The boxes that correspond to *TPS RML* model in Figure C.5 were computed using the information in these plots. These credible intervals have under coverage with respect to the nominal value 95% when the true variance of the errors is $\sigma^2 = .1^2$ and m is not small.

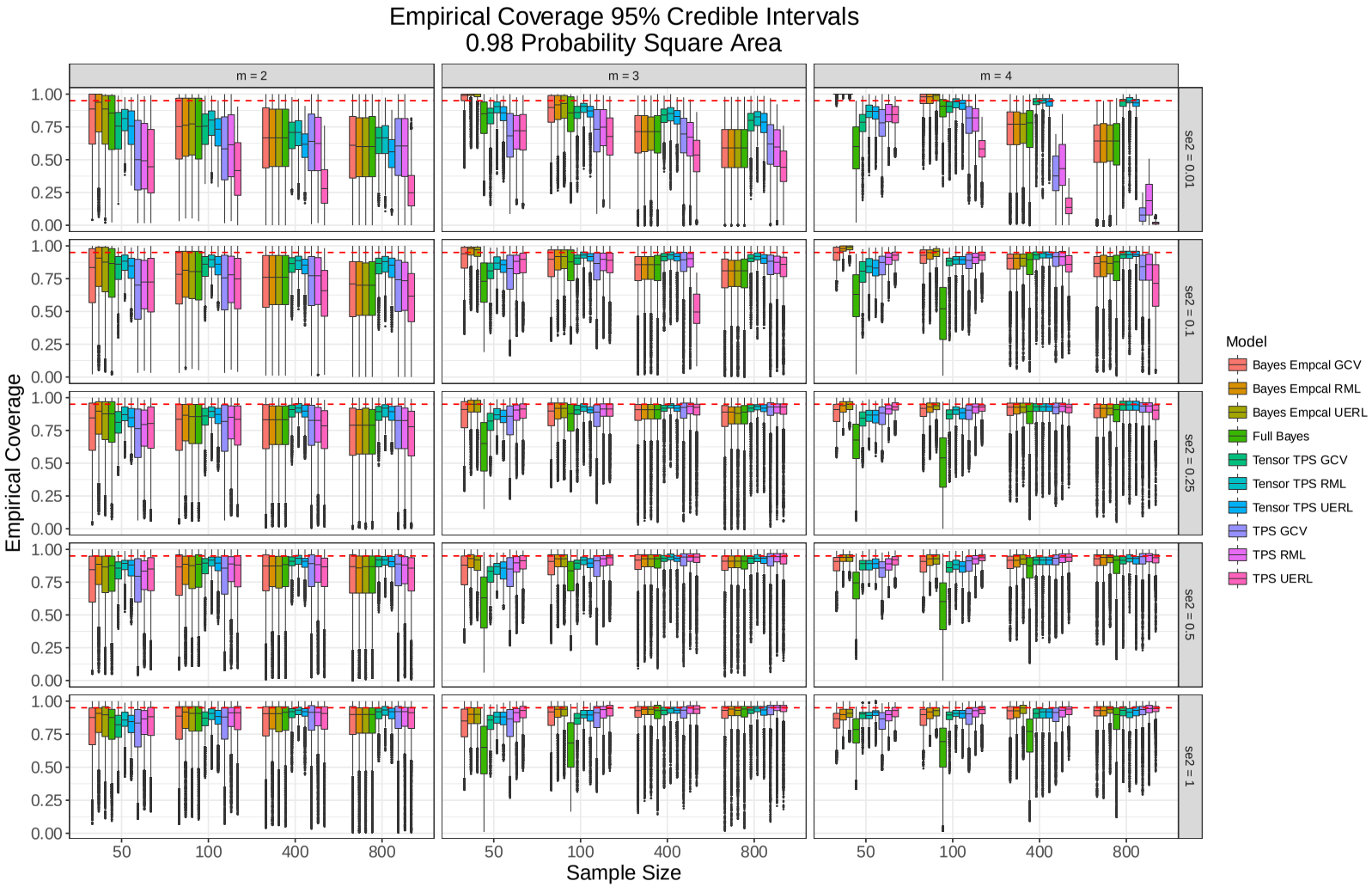

Figure C.5 Boxplots simulation results, empirical coverage of pointwise 95% credible intervals for prediction of multivariate regression functions. Each box is the summary of the empirical coverages $\{\hat{\rho}_i\}_{i=1}^{N}$. $\hat{\rho}_i \in [0,1]$ is the empirical coverage of the 95% pointwise credible interval for the prediction of $\eta(\chi_i)$ computed after fitting the model to 200 different simulated data sets. The vectors $\{\chi_i\}_{i=1}^{N} \subset \mathbb{R}^2$ is a grid of resolution 0.05×0.05 in the square $[-2.5, 2.5]^2$.

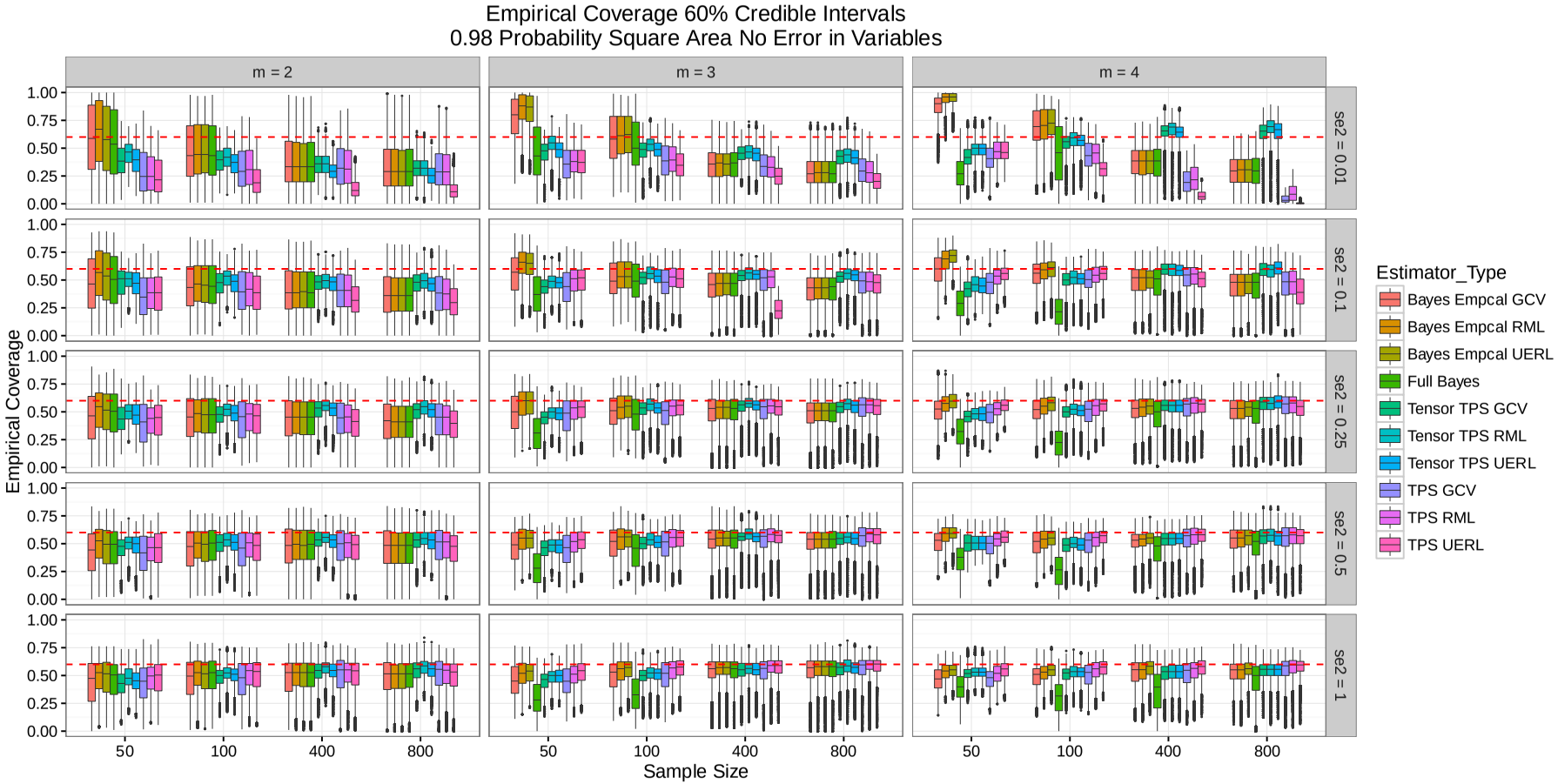

Figure C.6 Boxplots simulation results, empirical coverage of pointwise 60% credible intervals for prediction of multivariate regression functions. Each box is the summary of the empirical coverages $\{\hat{\rho}_i\}_{i=1}^{N}$. $\hat{\rho}_i \in [0,1]$ is the empirical coverage of the 95% pointwise credible interval for the prediction of $\eta(\chi_i)$ computed after fitting the model to 200 different simulated data sets. The vectors $\{\chi_i\}_{i=1}^{N} \subset \mathbb{R}^2$ is a grid of resolution 0.05×0.05 in the square $[-2.5, 2.5]^2$.

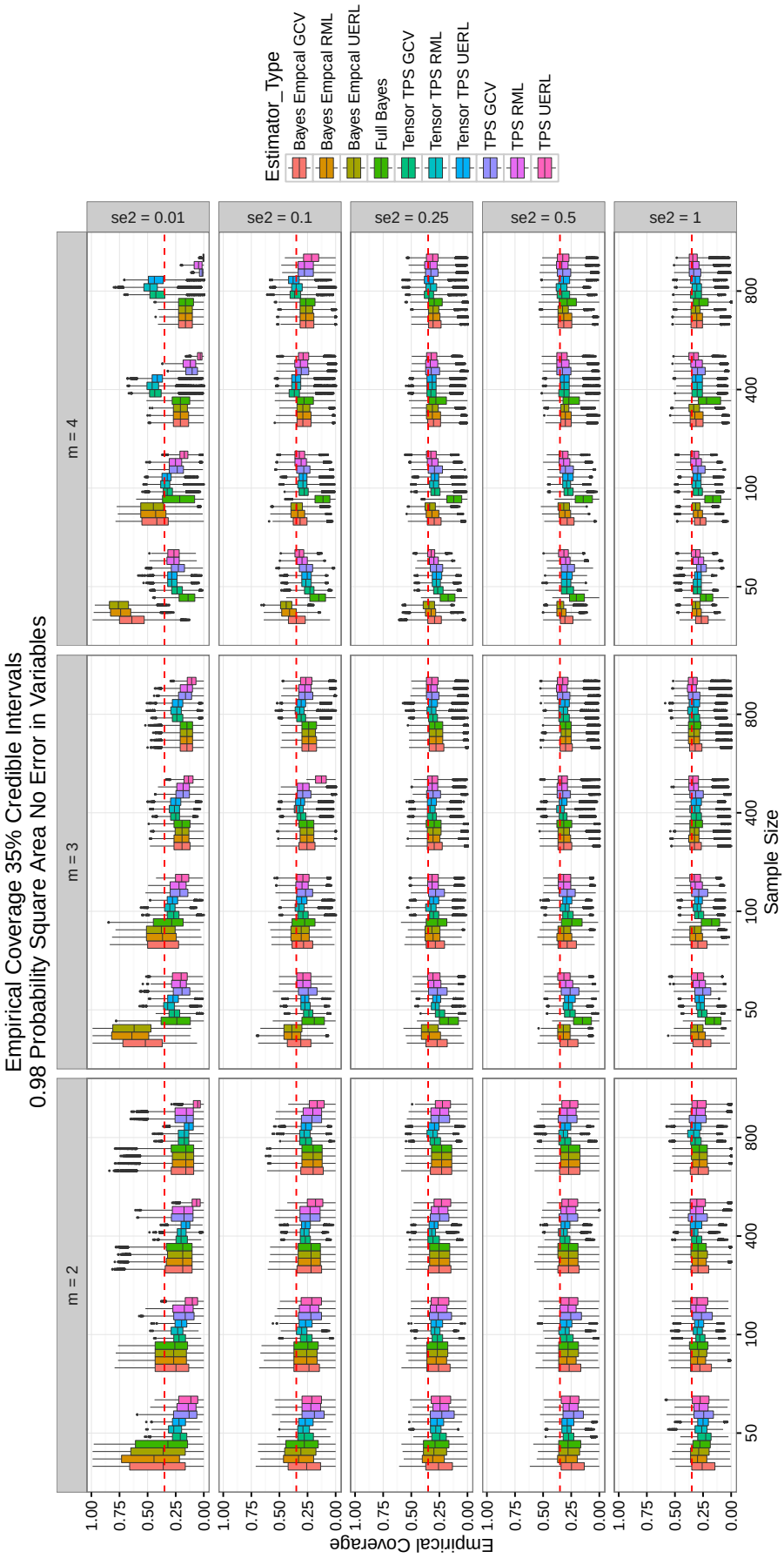

Figure C.7 Boxplots simulation results, empirical coverage of pointwise 35% credible intervals for prediction of multivariate regression functions. Each box is the summary of the empirical coverages $\{\hat{\rho}_i\}_{i=1}^N$. $\hat{\rho}_i \in [0,1]$ is the empirical coverage of the 95% pointwise credible interval for the prediction of $\eta(\chi_i)$ computed after fitting the model to 200 different simulated data sets. The vectors $\{\chi_i\}_{i=1}^N \subset \mathbb{R}^2$ is a grid of resolution 0.05×0.05 in the square $[-2.5, 2.5]^2$.

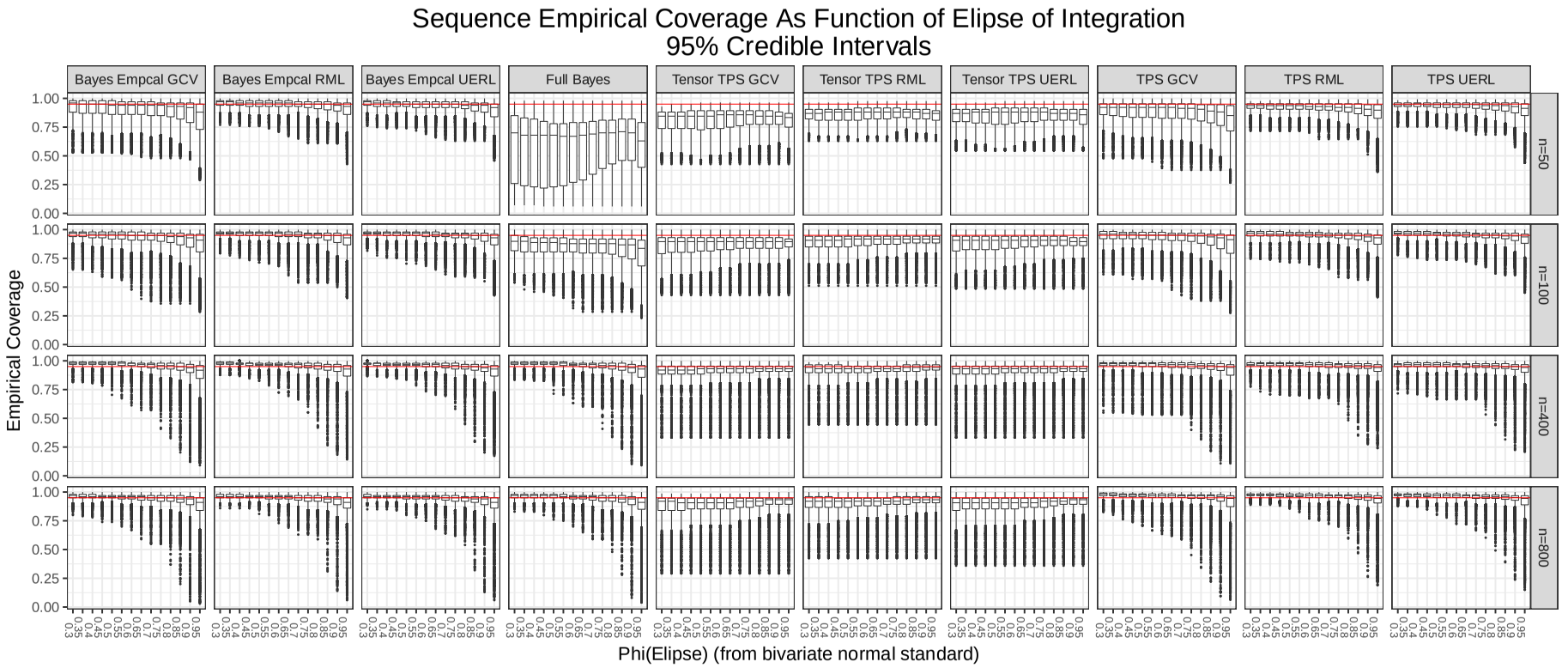

Figure C.8 Boxplots simulation results; sequential empirical coverage of pointwise 95% credible intervals for prediction of multivariate regression function I. In the $x - axis$, *Ellipse* is the ellipse region in $\mathbb{R}^2$ that would contain about $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution; $\alpha = Phi(Ellipse)$. Each box is the the summary of the empirical coverages $\hat{\rho}_i$'s for 95% pointwise credible intervals for the values of $\eta(\chi_i)$, and χ_i inside the ellipse region. The sequence is in the sense of observing the summaries of the $\hat{\rho}_i$'s as α changes. The horizontal red line has a value in the vertical axis of 0.95; the nominal value of the credible intervals is 95%. Here the simulated data was obtained using $\sigma^2 = 0.5$ and the models were fitted using $m = 3$. The plot is not a comprehensive summary of the simulation study.

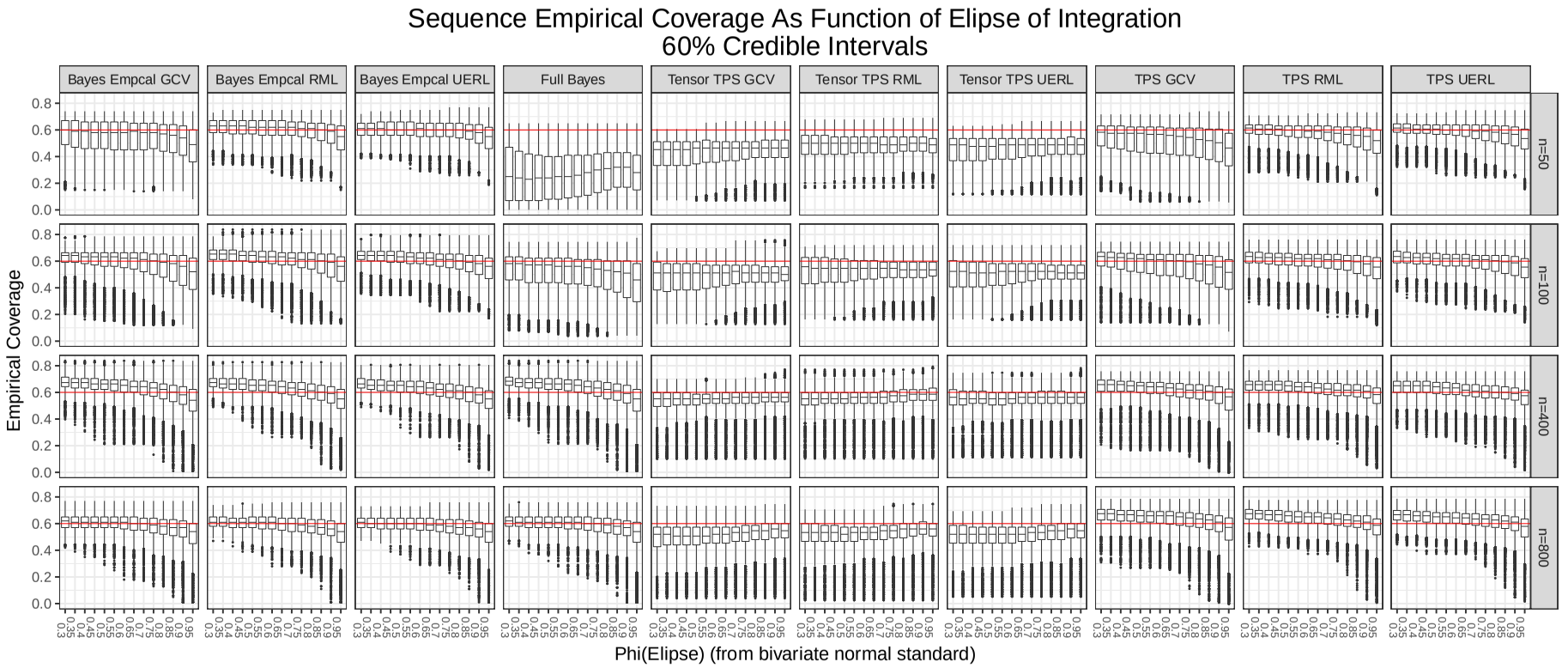

Figure C.9 Boxplots simulation results; sequential empirical coverage of pointwise 60% credible intervals for prediction of multivariate regression function I. In the $x - axis$, *Ellipse* is the ellipse region in $\mathbb{R}^2$ that would contain about $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution; $\alpha = Phi(Ellipse)$. Each box is the the summary of the empirical coverages $\hat{\rho}_i$'s for 60% pointwise credible intervals for the values of $\eta(\chi_i)$, and χ_i inside the ellipse region. The sequence is in the sense of observing the summaries of the $\hat{\rho}_i$'s as α changes. The horizontal red line has a value in the vertical axis of 0.60; the nominal value of the credible intervals is 60%. Here the simulated data was obtained using $\sigma^2 = 0.5$ and the models were fitted using $m = 3$. The plot is not a comprehensive summary of the simulation study.

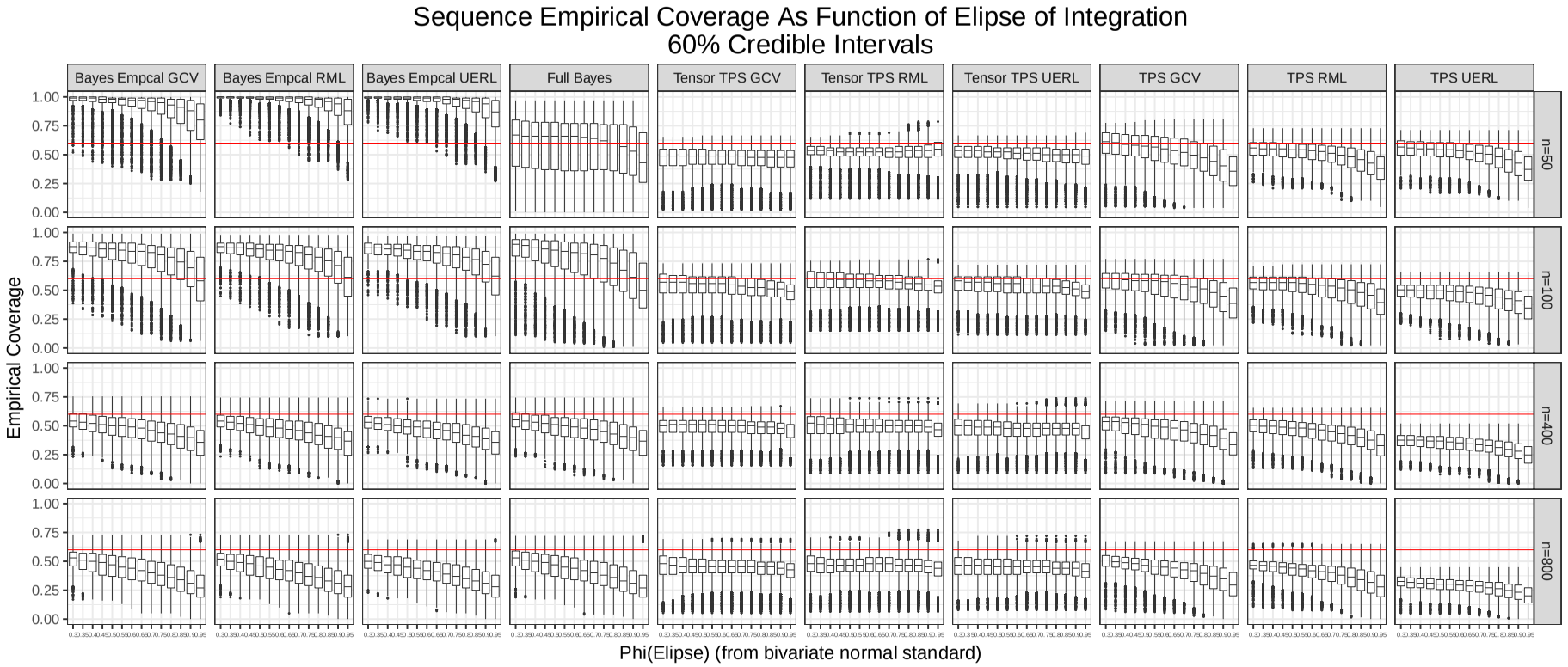

Figure C.10 Boxplots simulation results; sequential empirical coverage of pointwise 60% credible intervals for prediction of multivariate regression function II. In the $x-axis$, *Ellipse* is the ellipse region in $\mathbb{R}^2$ that would contain about $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution; $\alpha = Phi(Ellipse)$. Each box is the the summary of the empirical coverages $\hat{\rho}_i$'s for 60% pointwise credible intervals for the values of $\eta(\chi_i)$, and χ_i inside the ellipse region. The sequence is in the sense of observing the summaries of the $\hat{\rho}_i$'s as α changes. The horizontal red line has a value in the vertical axis of 0.60; the nominal value of the credible intervals is 60%. Here the simulated data was obtained using $\sigma^2 = 0.1^2$ and the models were fitted using $m = 3$. The plot is not a comprehensive summary of the simulation study.

C.2 Real Valued Functions Regression with Measurement Errors in the Covariates

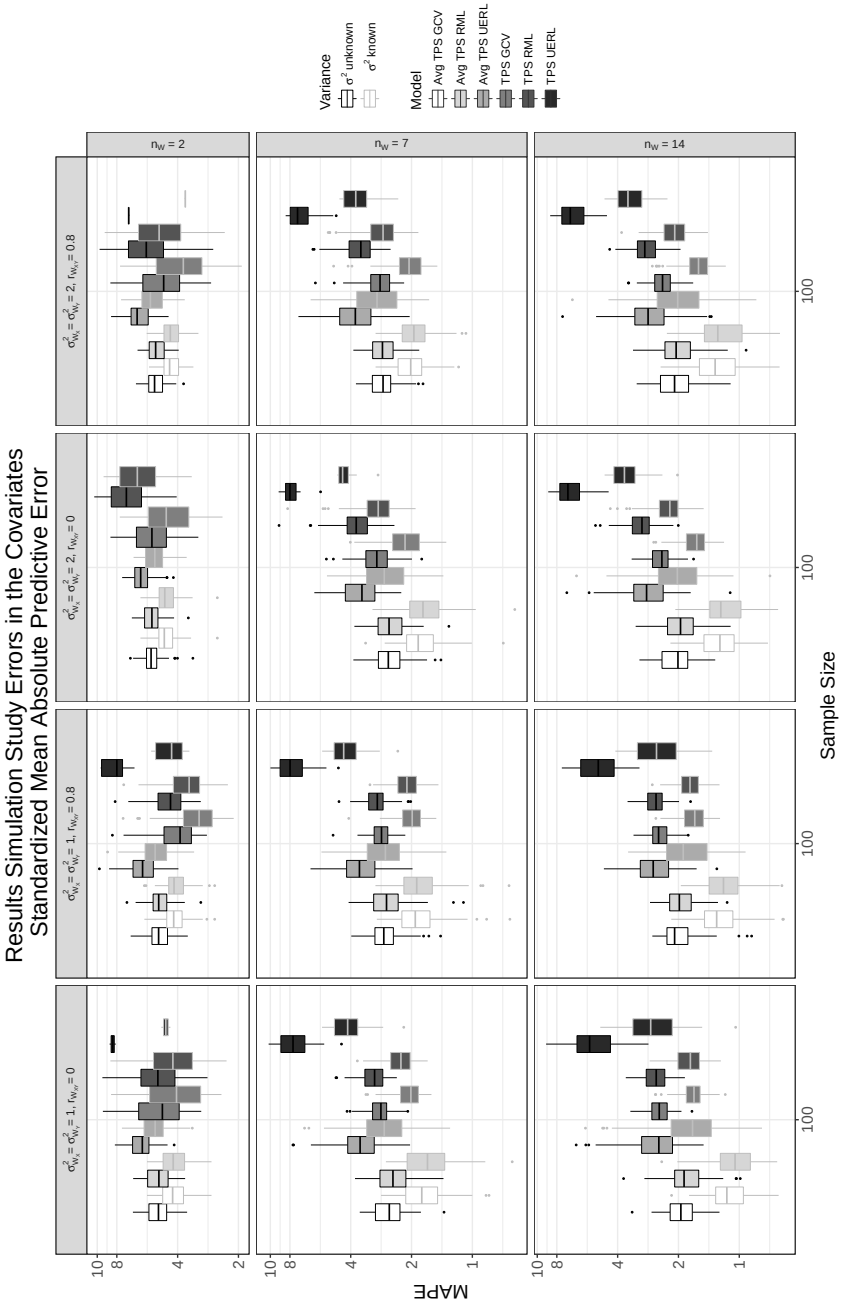

Figure C.11 Box plots simulation results. Standardized Mean Absolute Predictive Error (3.19) for the multivariate regression problem with measurement errors in the covariates. The MAPE was computed over the square $[-2.25, 2.25]^2$. This is the full graphical display of the simulation study from table 4.2. The simulated data sets $\{(\mathbf{w}_{i1}, \cdots \mathbf{w}_{in_w}, y_i)\}_{i=1}^{n}$ were generated using the model (4.1), $n = 100$ and $\sigma^2 = 0.5$. The columns in the figure indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the number of repeated observations of the measurement errors of the covariates $\{\mathbf{w}_{j,i}\}_{i=1}^{n_w}$. Observe that the $y - axis$ is in the log_{10} scale. The models are described in Table 4.1.

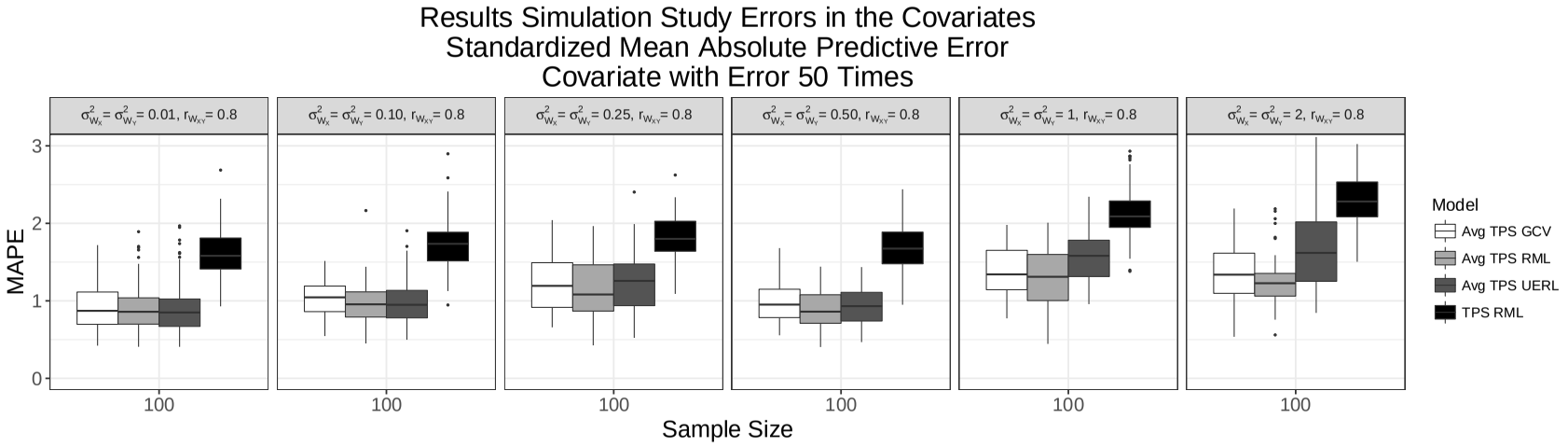

Figure C.12 Box plots simulation results. Standardized Mean Absolute Predictive Error (3.19) for the multivariate regression problem with measurement errors in the covariates, $n_w = 50$. The MAPE was computed over the square $[-2.25, 2.25]^2$. The simulated data sets $\{(\mathbf{w}_{i1}, \cdots \mathbf{w}_{i50}, y_i)\}_{i=1}^{n}$ were generated using the model (4.1), $n = 100$ and $\sigma^2 = 0.5$. The columns in the figure indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the number of repeated observations of the measurement covariates $\{\mathbf{x}_i\}_{i=1}^{n_w}$ with errors. The models are described in Table 4.1.

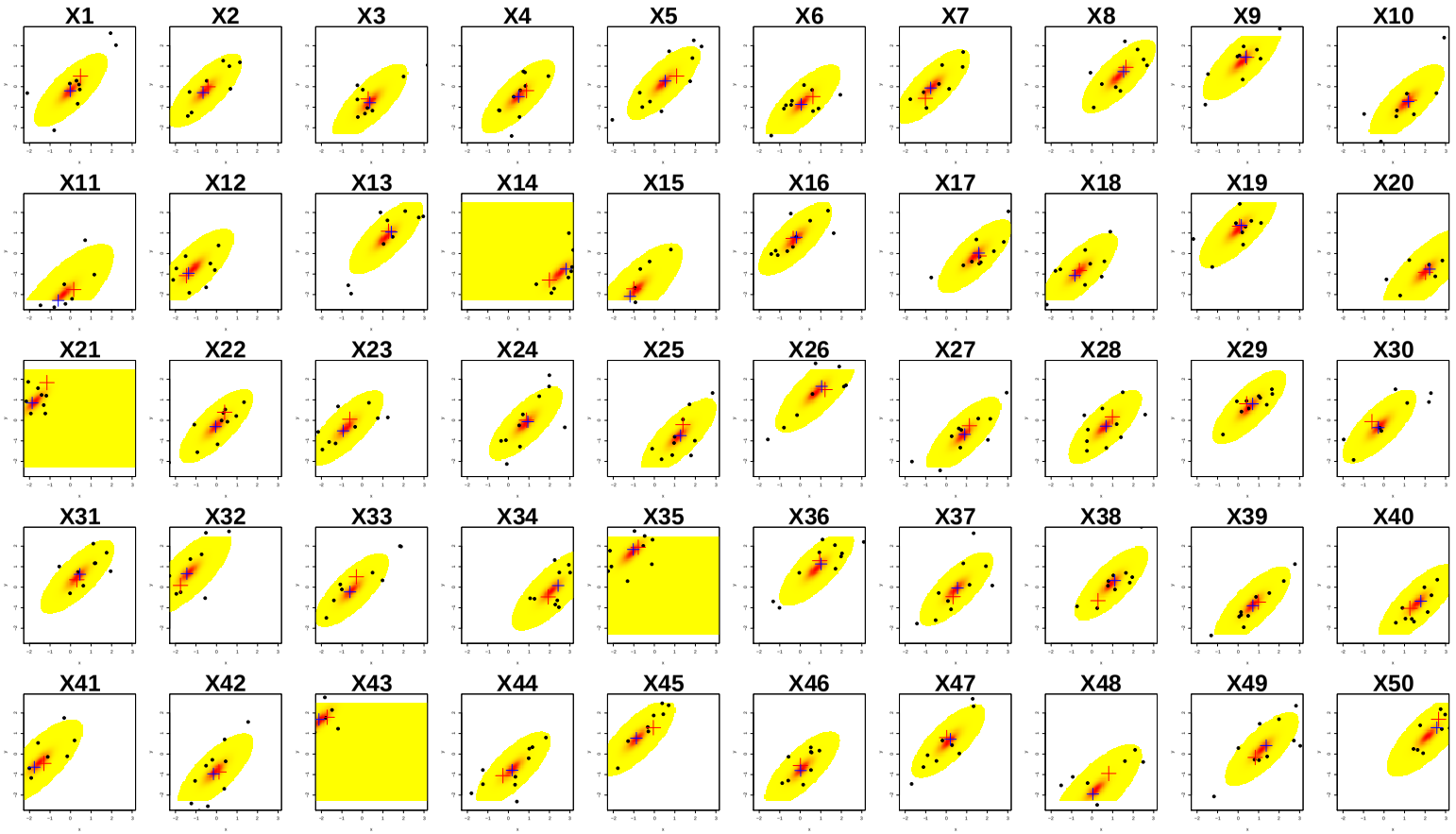

Figure C.13 Level curves of posterior density of latent variables; an example. The data sets $\{(\mathbf{w}_{i1}, \cdots \mathbf{w}_{i7}, y_i)\}_{i=1}^{50}$ were simulated using the form (4.1), with $\sigma^2 = 0.5^2$, $\mathbf{\Sigma}_w = \begin{pmatrix} 2 & 0.8\times 2 \\ 0.8\times 2 & 2 \end{pmatrix}$, and estimated using a conditional thin plate spline, $m = 3$ model with inverse gamma prior on σ^2. These plots show the estimate of the latent variables $\{\mathbf{x}_i\}_{i=1}^{50}$. The dots in each plot are the measurements with errors, the red cross is the true latent variable, the blue cross is the location of the average of the measures with errors, the yellow region indicates the limits of 95% credible region while the red region is approximately the 50% centered credible region obtained from the marginal posterior of $\mathbf{x}_i$.

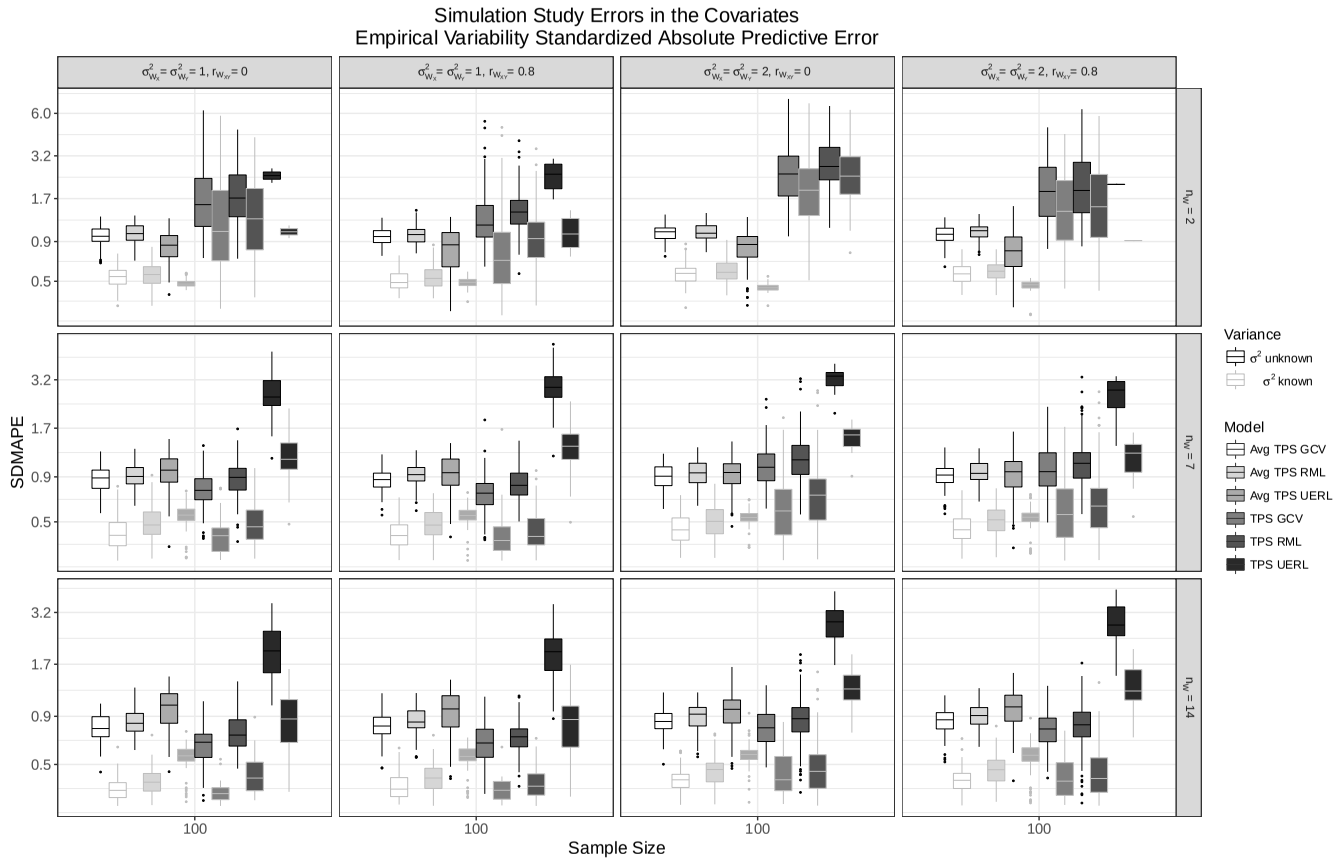

Figure C.14 Variability Standardized Absolute Predictive Error (3.20) for the multivariate regression problem with measurement errors in the covariates. The SDMAPE was computed over the square $[-2.25, 2.25]^2$. This is the full graphical display of the simulation study from table 4.2, last three columns. The simulated data sets $\{(\mathbf{w}_{i1}, \cdots \mathbf{w}_{in_w}, y_i)\}_{i=1}^{100}$ were generated using the model (4.1) and $\sigma^2 = 0.5$. The columns in the figure indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the number of repeated observations of the measurement covariates $\{\mathbf{x}_i\}_{i=1}^{n_w}$ with errors. Observe that the y-axis is in the log_{10} scale. The models are described in Table 4.1.

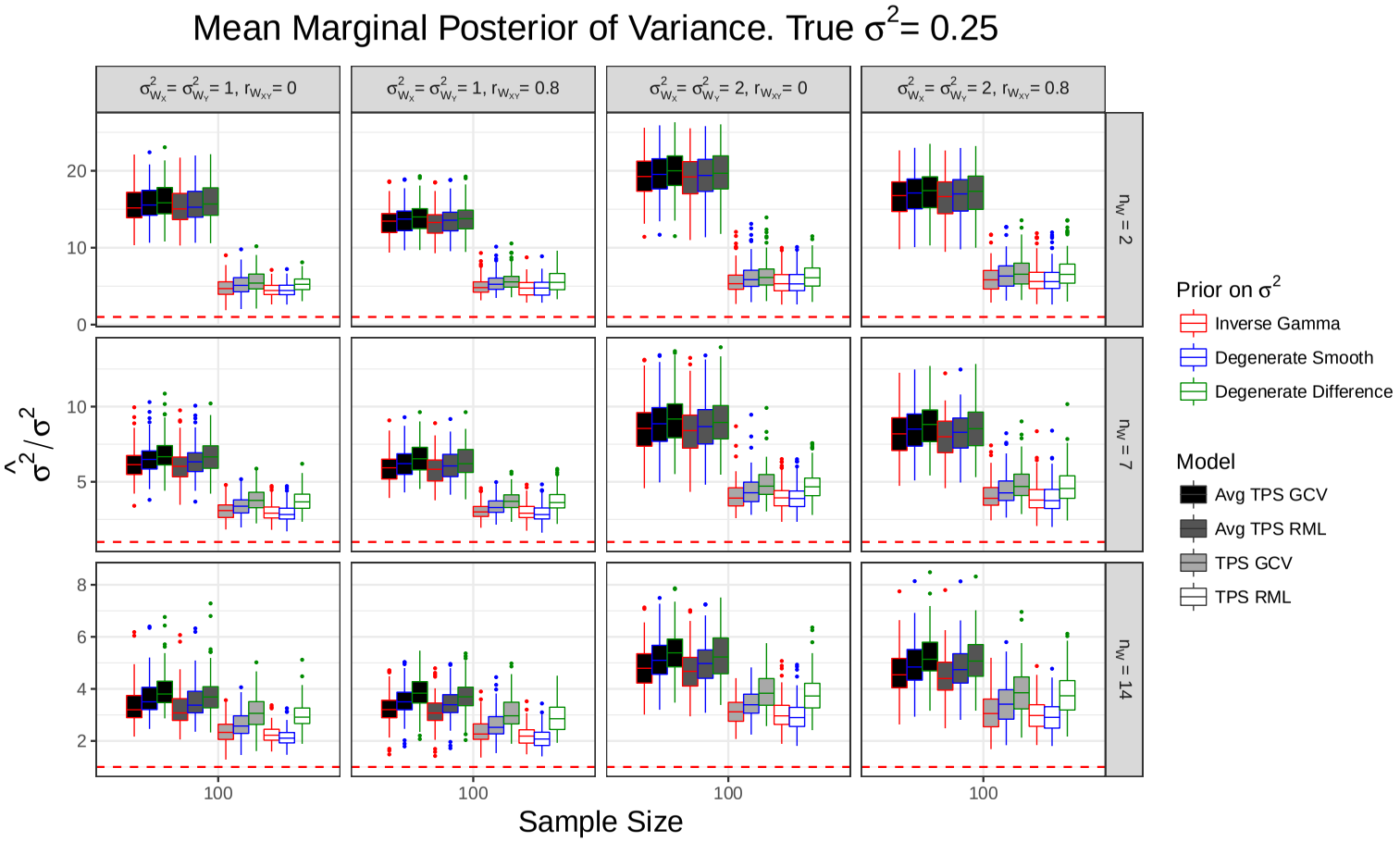

Figure C.15 Estimator for the Observation-error Variance σ^2 in the Multivariate Regression Problem with Measurement Errors in the Covariates I. The simulated data sets $\{(\mathbf{w}_{i1}, \cdots \mathbf{w}_{in_w}, y_i)\}_{i=1}^{100}$ were generated using the model (4.1) and $\sigma^2 = 0.5^2$. The columns in the figure indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the number of repeated observations of the measurement covariates $\{\mathbf{x}_i\}_{i=1}^{n_w}$ with errors. Observe that each plot has a different range of values. The models are described in Table 4.1.

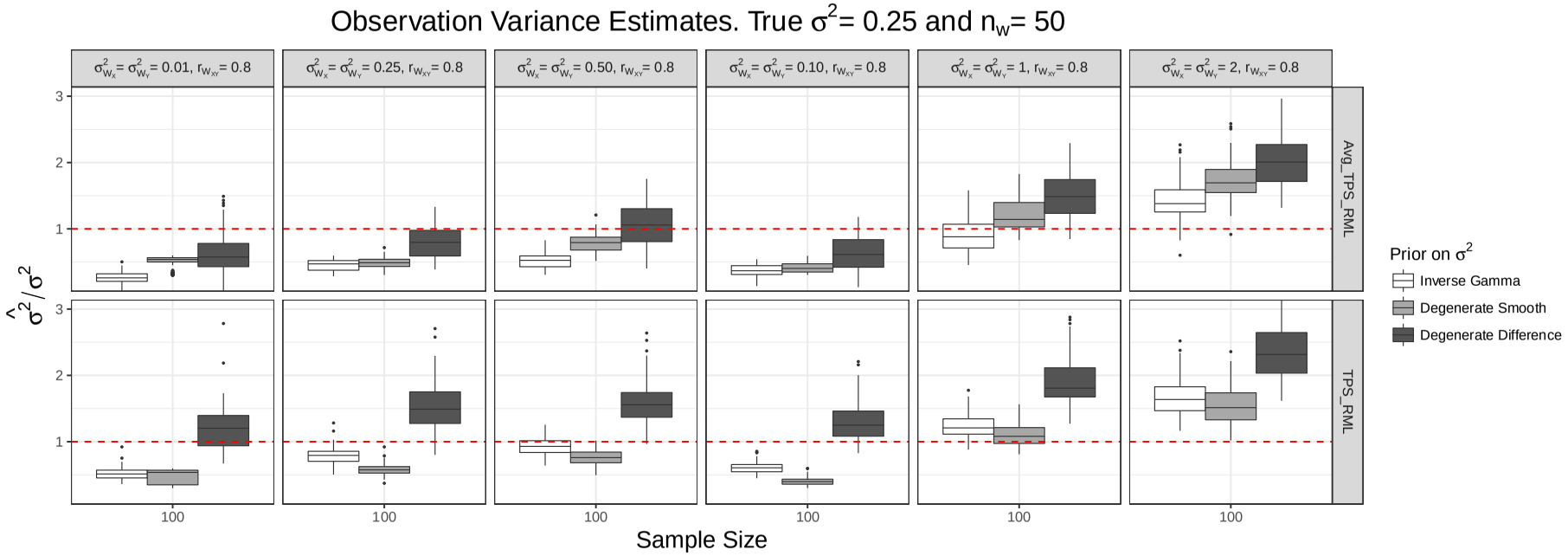

Figure C.16 Estimator for the Observation-error Variance σ^2 in the Multivariate Regression Problem with Measurement Errors in the Covariates II. The simulated data sets $\{(\mathbf{w}_{i,1}, \cdots \mathbf{w}_{i,50}, y_i)\}_{i=1}^{100}$ were generated using the model (4.1) and $\sigma^2 = 0.5^2$. The columns in the figure indicate the covariance matrix of the measurement errors $\left\{\{\delta_{ji}\}_{i=1}^{50}\right\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the model used to fit the data. The models are described in Table 4.1.

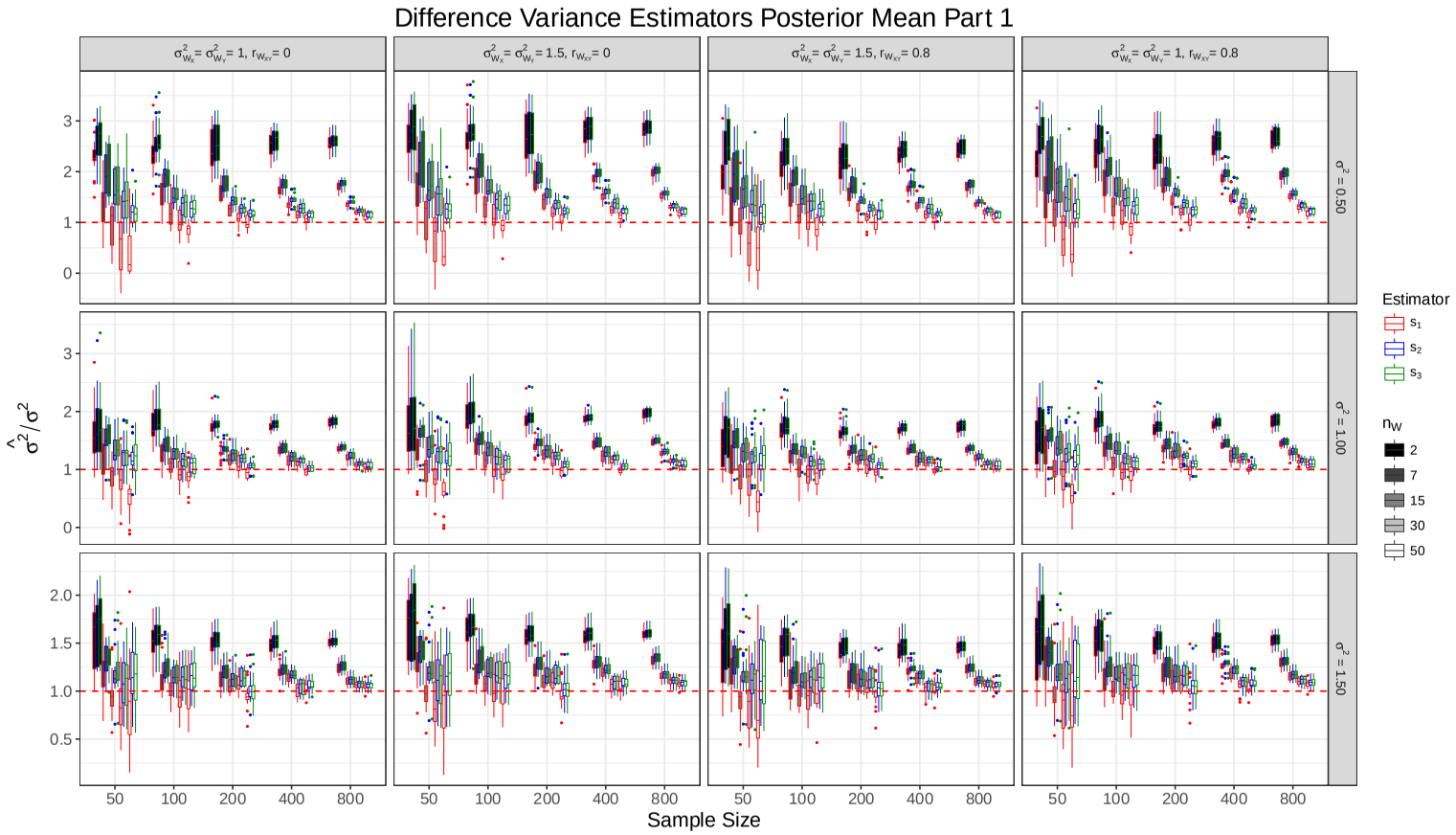

Figure C.17 Mean marginal posterior of the observation-error variance σ^2 in the multivariate regression problem with measurement errors in the covariates using difference method for the priors in the Bayes model I. The simulated data sets $\{(\mathbf{w}_{i,1}, \cdots \mathbf{w}_{i,n_w}, y_i)\}_{i=1}^{100}$ were generated using the model (4.1). The columns in the plot indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the true variance σ^2 to be estimated. The estimators s_1, s_2 and s_3 indicate that the expression (4.11), (4.12) and (4.13) respectively were used as prior for σ^2. Observe that the $y-axis$ is the ratio of the estimated variance by the true variance, and each plot has a different range of values in the vertical axis.

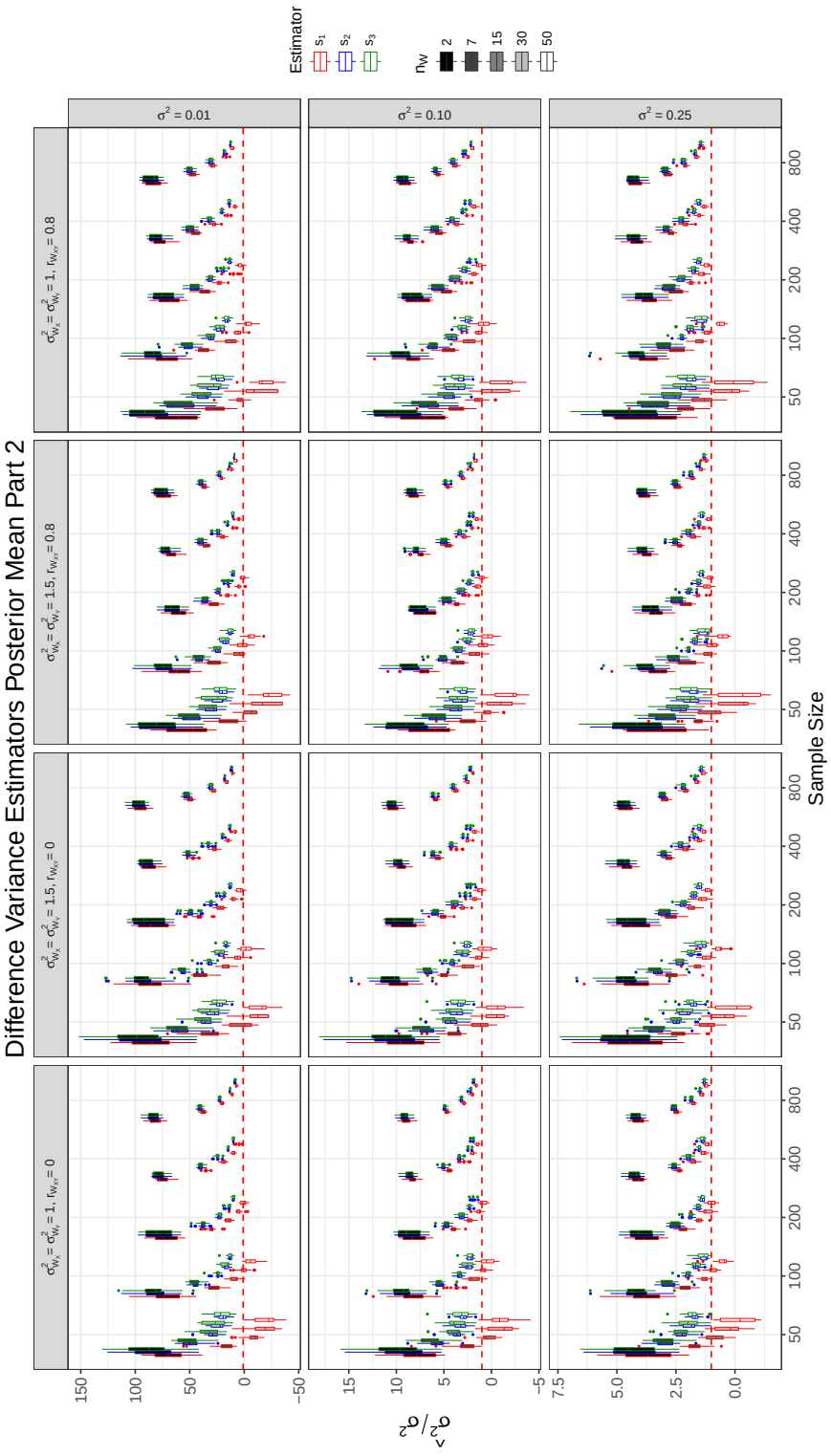

Figure C.18 Mean marginal posterior of the observation-error variance σ^2 in the multivariate regression problem with measurement errors in the covariates using difference method for the priors in the Bayes model II. The simulated data sets $\{(\mathbf{w}_{i,1}, \cdots \mathbf{w}_{i,n_w}, y_i)\}_{i=1}^{100}$ were generated using the model (4.1). The columns in the figure indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the true variance σ^2 to be estimated. The estimators s_1, s_2 and s_3 indicate that the expression (4.11), (4.12) and (4.13) respectively were used as prior for σ^2 in different models. Observe that the $y-axis$ is the ratio of the estimated variance by the true variance, and each plot has a different range of values in the vertical axis.

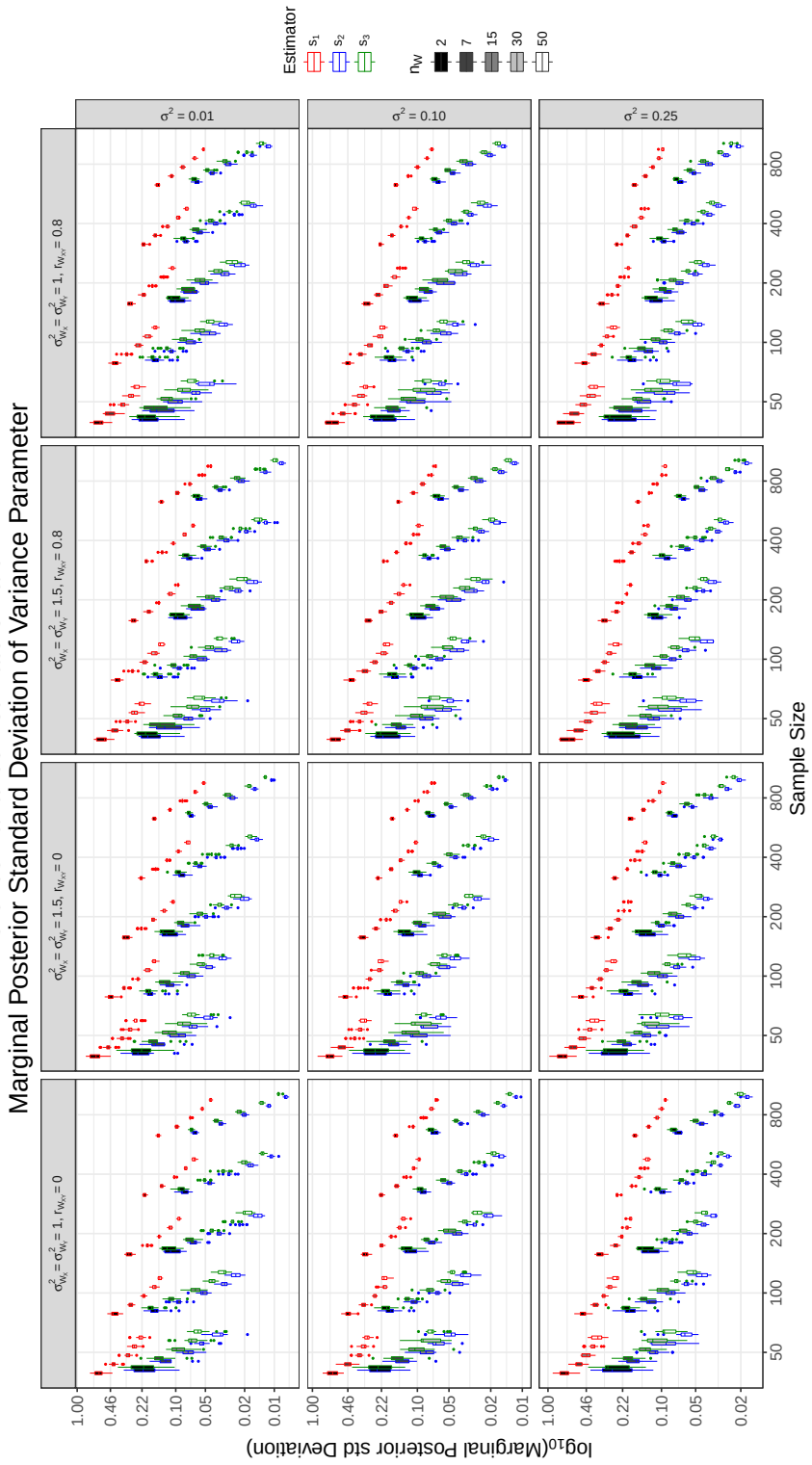

Figure C.19 Variance marginal posterior of the observation-error variance, σ^2, in the multivariate regression problem with measurement errors in the covariates using difference method for the priors in the Bayes model I. The simulated data sets $\{(\mathbf{w}_{i,1}, \cdots \mathbf{w}_{i,n_w}, y_i)\}_{i=1}^{100}$ were generated using the model (4.1). The columns in the plot indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the true variance σ^2 to be estimated. The estimators s_1, s_2 and s_3 indicate that the expression (4.11), (4.12) and (4.13) respectively were used as prior for σ^2 in different models. Observe that the $y-axis$ is the ratio of the estimated variance by the true variance, and each plot has a different range of values in the vertical axis.

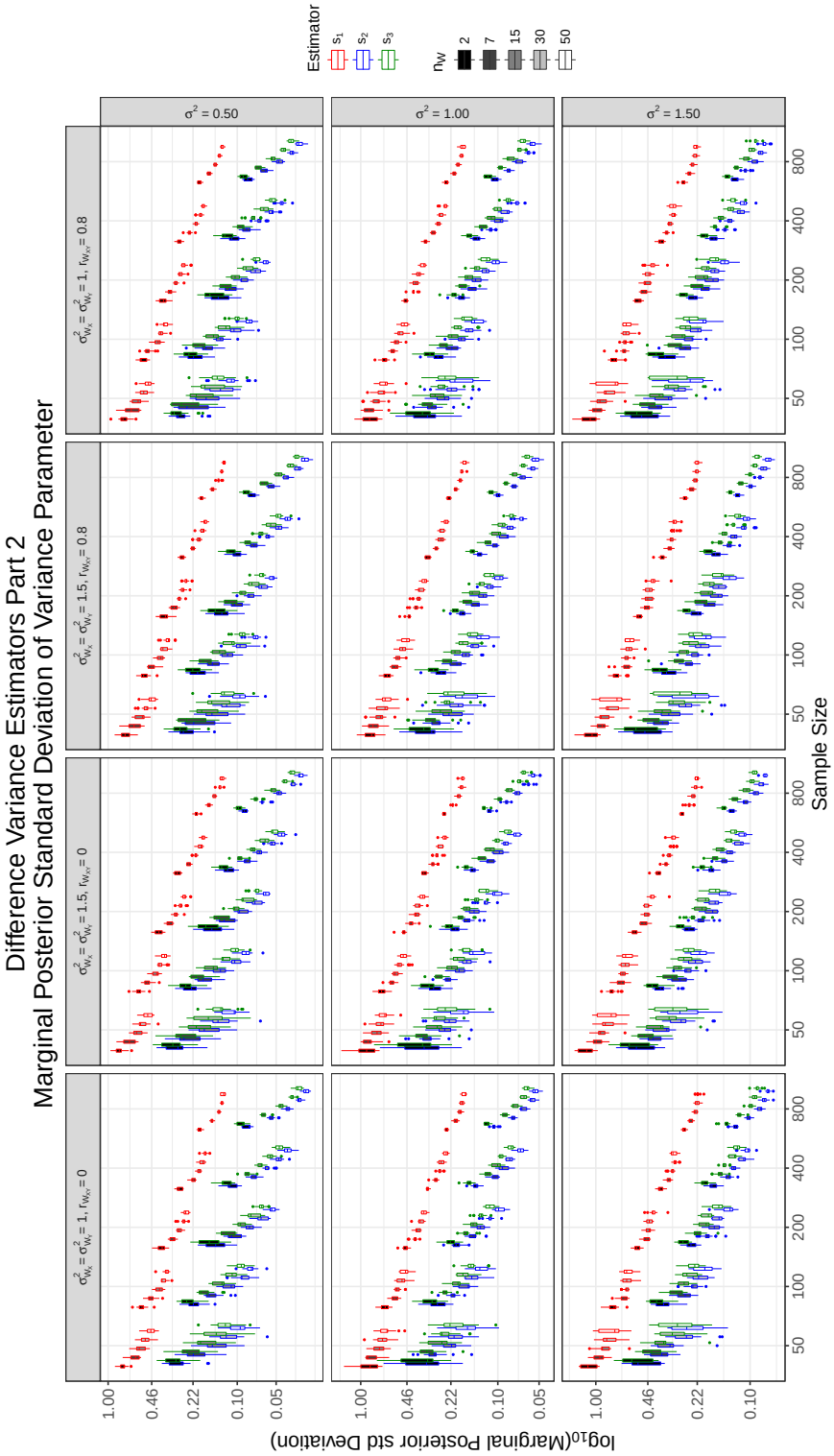

Figure C.20 Variance marginal posterior of the observation-error variance, σ^2, in the multivariate regression problem with measurement errors in the covariates using difference method for the priors in the Bayes model II. The simulated data sets $\{(\mathbf{w}_{i,1}, \cdots \mathbf{w}_{i,n_w}, y_i)\}_{i=1}^{100}$ were generated using the model (4.1). The columns in the plot indicate the covariance matrix of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated, and the rows indicate the true variance σ^2 to be estimated. The estimators s_1, s_2 and s_3 indicate that the expression (4.11), (4.12) and (4.13) respectively were used as prior for σ^2 in different models. Observe that the $y-axis$ is the ratio of the estimated variance by the true variance, and each plot has a different range of values in the vertical axis.

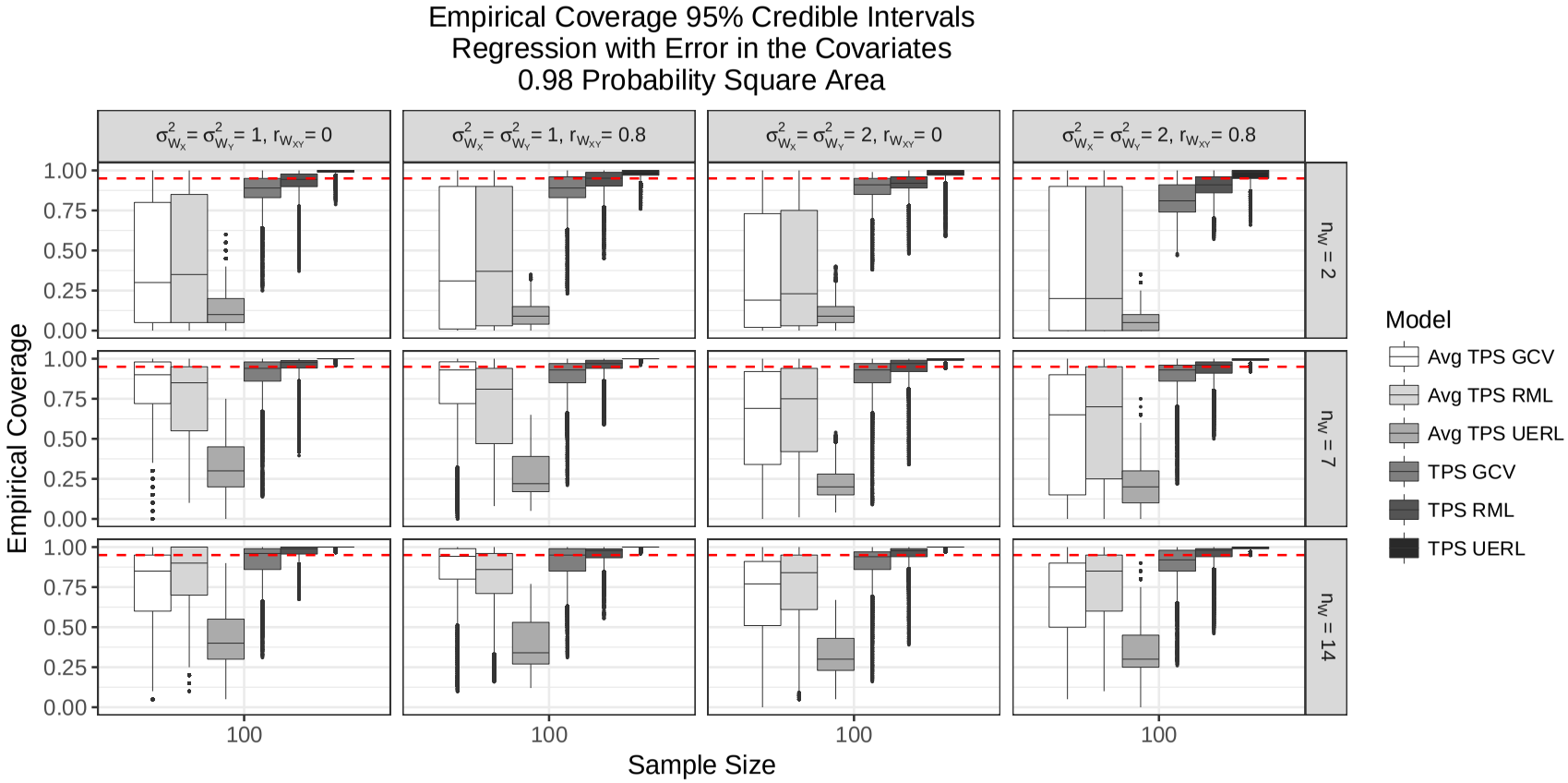

Figure C.21 Boxplots simulation results, empirical coverage of pointwise 95% credible intervals for prediction of multivariate regression functions with measurement errors in the covariates I. Each box is the summary of the empirical coverages $\{\hat{\rho}_i\}_{i=1}^{N}$. $\hat{\rho}_i \in [0,1]$ is the empirical coverage of the 95% pointwise credible interval for the prediction of $\eta(\chi_i)$ computed after fitting the model to 150 different simulated data sets. The vectors $\{\chi_i\}_{i=1}^{N} \subset \mathbb{R}^2$ is a grid of resolution 0.05×0.05 in the square $[-2.5, 2.5]^2$. The columns in the plot indicate the covariance matrix $\Sigma_W = \begin{pmatrix} \sigma^2_{W_X} & \sigma_{W_X}\sigma_{W_Y} r_{W_{XY}} \\ \sigma_{W_X}\sigma_{W_Y} r_{W_{XY}} & \sigma^2_{W_Y} \end{pmatrix}$ of the measurement errors $\{\{\delta_{ji}\}_{i=1}^{n_w}\}_{j=1}^{100}$ from which the data were generated. The models are described in Table 4.1.

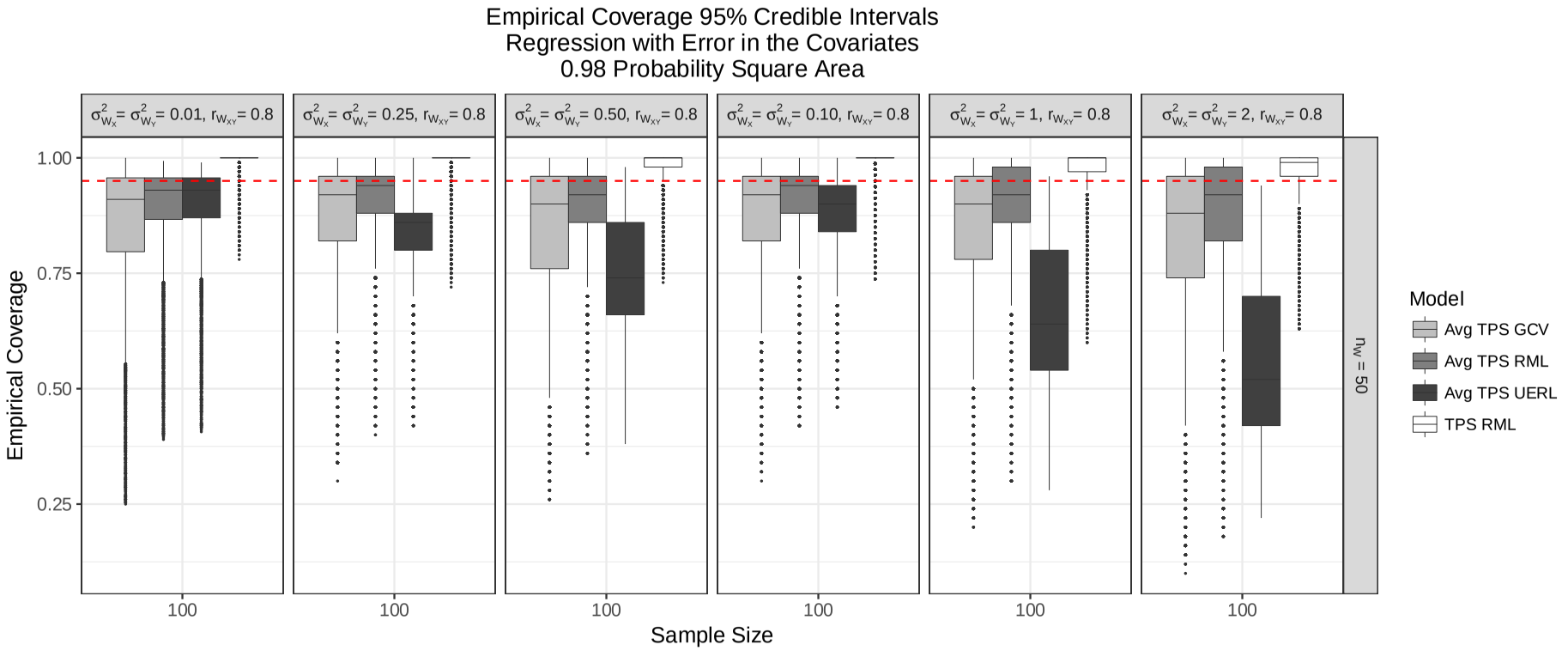

Figure C.22 Boxplots simulation results, empirical coverage of pointwise 95% credible intervals for prediction of multivariate regression functions with measurement errors in the covariates I. Each box is the summary of the empirical coverages $\{\hat{\rho}_i\}_{i=1}^{N}$. $\hat{\rho}_i \in [0,1]$ is the empirical coverage of the 95% pointwise credible interval for the prediction of $\eta(\chi_i)$ computed after fitting the model to 150 different simulated data sets. The vectors $\{\chi_i\}_{i=1}^{N} \subset \mathbb{R}^2$ is a grid of resolution 0.05×0.05 in the square $[-2.5, 2.5]^2$. The columns in the plot indicate the covariance matrix $\Sigma_W = \begin{pmatrix} \sigma^2_{W_X} & \sigma_{W_X}\sigma_{W_Y} r_{W_{XY}} \\ \sigma_{W_X}\sigma_{W_Y} r_{W_{XY}} & \sigma^2_{W_Y} \end{pmatrix}$ of the measurement errors $\left\{\{\delta_{ji}\}_{i=1}^{50}\right\}_{j=1}^{100}$ from which the data were generated. The models are described in Table 4.1.

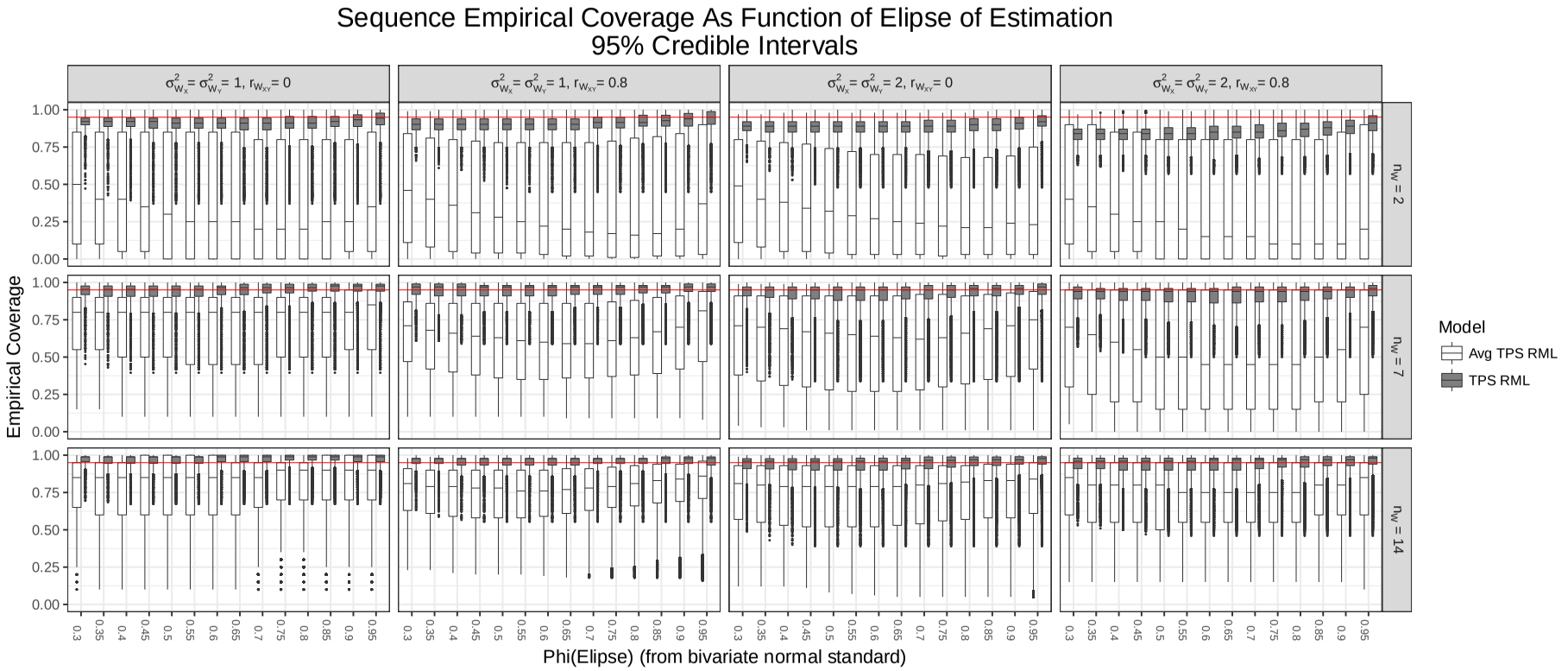

Figure C.23 Sequential empirical coverage of pointwise 95% credible intervals. The sequence is in the the sense of summarizing the pointwise empirical coverage of the credible intervals $\{\xi_i\}_{i=1}^{m_\alpha}$ for η evaluated in the points $\{\chi_i\}_{i=1}^{m_\alpha}$ from the grid and inside the ellipse that would contain $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution. $alpha = Phi(Ellipse)$. The 150 data sets used to fit the models and compute the empirical coverage were simulated using $n = 100$, $\sigma^2 = 0.25$, $\mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$ and $\mathbf{w}_{ij} \overset{iid}{\sim} N_2\left(\mathbf{x}_i, \begin{pmatrix} \sigma^2_{W_X} & \sigma_{W_X}\sigma_{W_Y} r_{W_{XY}} \\ \sigma_{W_X}\sigma_{W_Y} r_{W_{XY}} & \sigma^2_{W_Y} \end{pmatrix}\right)$.

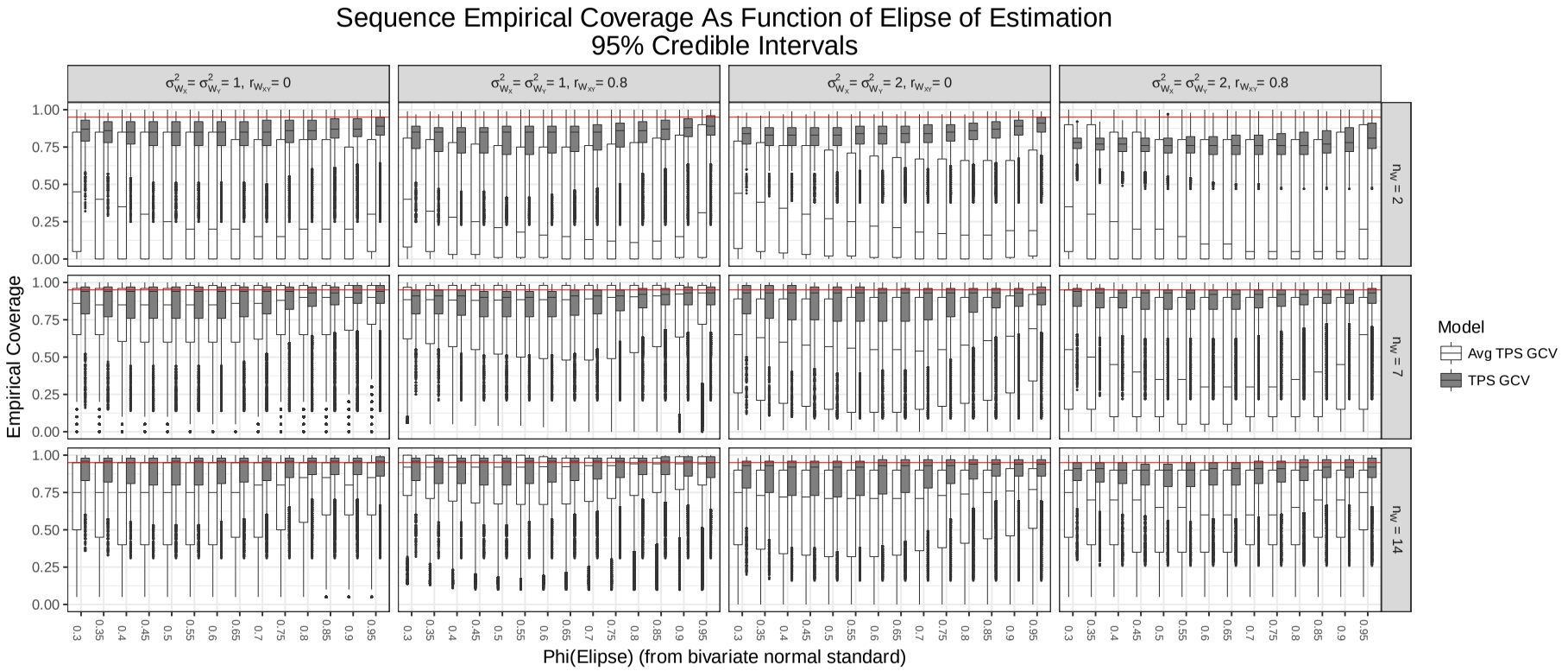

Figure C.24 Sequential empirical coverage of pointwise 95% credible intervals. The sequence is in the the sense of summarizing the pointwise empirical coverage of the credible intervals $\{\xi_i\}_{i=1}^{m_\alpha}$ for η evaluated in the points $\{\chi_i\}_{i=1}^{m_\alpha}$ from the grid and inside the ellipse that would contain $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution. $alpha = Phi(Ellipse)$. The 100 data sets used to fit the models and compute the empirical coverage were simulated using $n = 100$, $\sigma^2 = 0.25$, $\mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$ and $\mathbf{w}_{ij} \overset{iid}{\sim} N_2\left(\mathbf{x}_i, \begin{pmatrix} \sigma^2_{W_X} & \sigma_{W_X}\sigma_{W_Y}r_{W_{XY}} \\ \sigma_{W_X}\sigma_{W_Y}r_{W_{XY}} & \sigma^2_{W_Y} \end{pmatrix}\right)$.

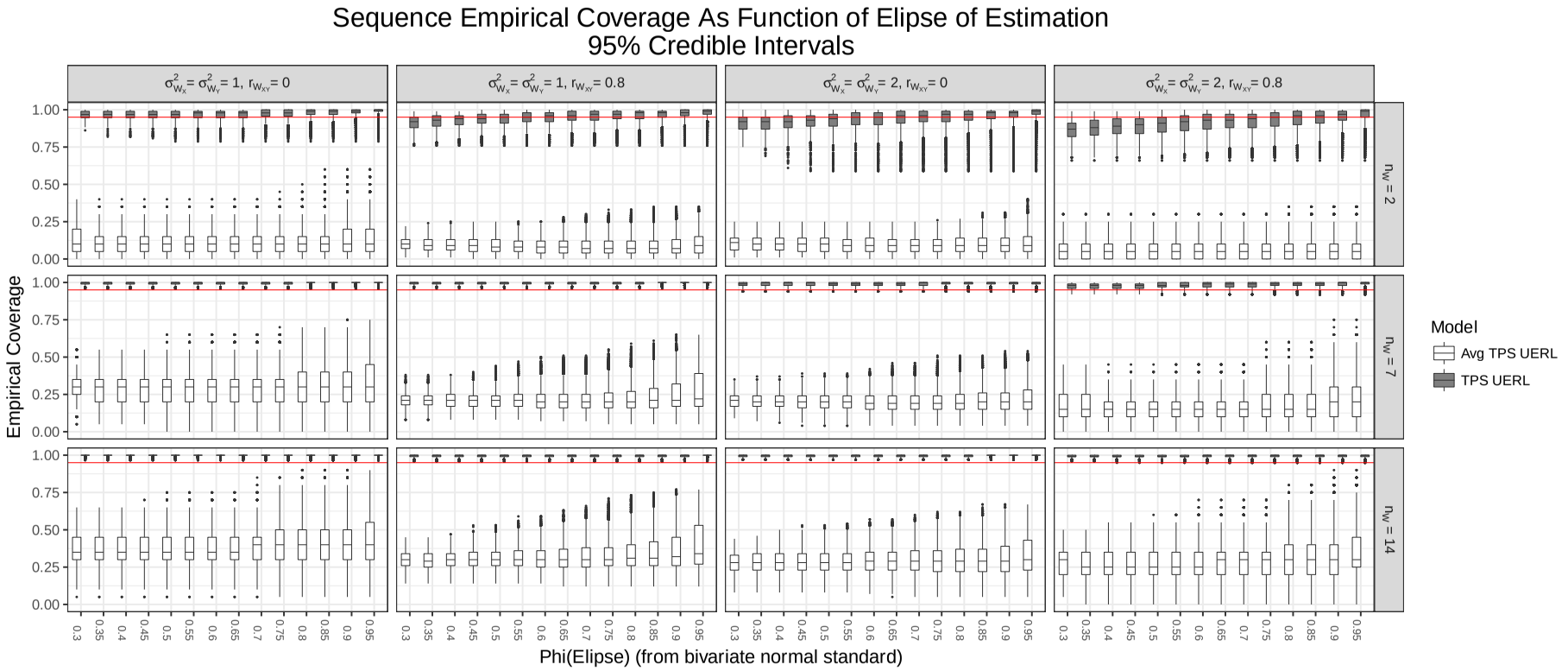

Figure C.25 Sequential empirical coverage of pointwise 95% credible intervals. The sequence is in the the sense of summarizing the pointwise empirical coverage of the credible intervals $\{\xi_i\}_{i=1}^{m_\alpha}$ for η evaluated in the points $\{\chi_i\}_{i=1}^{m_\alpha}$ from the grid and inside the ellipse that would contain $\alpha \times 100\%$ of the points generated from a standard bivariate normal distribution. $alpha = Phi(Ellipse)$. The 150 data sets used to fit the models and compute the empirical coverage were simulated using $n = 100$, $\sigma^2 = 0.25$, $\mathbf{x}_i \overset{iid}{\sim} N_2(\mathbf{0}, I_2)$ and $\mathbf{w}_{ij} \overset{iid}{\sim} N_2\left(\mathbf{x}_i, \begin{pmatrix} \sigma^2_{W_X} & \sigma_{W_X}\sigma_{W_Y}r_{W_{XY}} \\ \sigma_{W_X}\sigma_{W_Y}r_{W_{XY}} & \sigma^2_{W_Y} \end{pmatrix}\right)$.

APPENDIX D. NUMERICALLY STABLE COMPUTATIONS

In this appendix, we describe the practical difficulties to compute matrices required for the evaluation of functions and inversion of matrices used in Chapters 2, 3 and 4. We developed our methods using alternatives expressions for the required matrices, but after sometime, these approaches were found in Kim and Gu (2004). We describe matrix expressions to compute the inversion of large matrices needed in algorithms to fit the models.

Let $\{\mathbf{x}_i, y_i\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ be a given training set, let be $\{\mathbf{z}_i\}_{i=1}^k \subset \{\mathbf{x}_i\}_{i=1}^n$ a set of knots, a reproducing kernel $R_J(\cdot,\cdot)$ in a *RKHS* as in chapter 2 – 4, a basis set of functions $\{\psi_i\}_{i=1}^l$ of the null space $\mathcal{N}_J$, $Q \in \mathcal{M}_{k\times k}(\mathbb{R})$ with entries $Q_{i,j} = R_J(\mathbf{z}_i, \mathbf{z}_j)$, $S \in \mathcal{M}_{n\times l}(\mathbb{R})$ with $S_{i,j} = \psi_j(\mathbf{x}_i)$ full column rank, $R \in \mathcal{M}_{n\times k}(\mathbb{R})$ full column matrix, $R_{i,j} = R_J(\mathbf{x}_i, \mathbf{z}_j)$ and $M = RQ^+R^\intercal + n\lambda I_n$.

We want to compute $\left(S^\intercal M^{-1} S\right)^{-1} \in \mathcal{M}_{l\times l}(\mathbb{R})$ and $M^{-1} \in \mathcal{M}_{n\times n}(\mathbb{R})$. These expressions are required in the evaluation of the functions $\mathcal{U}$ (2.55), $\mathcal{V}$ (2.59), $\mathcal{M}$ (2.72) and they appear in the full conditional posterior of all our models. Inversion of the $n \times n$ matrices can be accomplished using standard procedures taking into account that the matrices are symmetric. Nevertheless, the computation of the inverses in this way leads to numerically unstable and slow evaluations of the scores $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{M}$. This is an important problem for the algorithms proposed to fit the Bayesian models, especially in Chapter 4 where each step of the metropolis within Gibbs algorithm requires to evaluate one of the scores, minimize such score, to compute M^{-1} and $(S^\intercal M^{-1} S)$. Figure D.1 shows the evaluation of the scores $\mathcal{U}$, $\mathcal{V}$ and $\mathcal{M}$ using their definition. The evaluation of these scores and the numerical instability is evident in the Restrictive Maximum Likelihood method $\mathcal{M}$.

It can be proven directly that

$$M^{-1} = \frac{1}{n\lambda}\left\{I - R\left(n\lambda Q + R^\intercal R\right)^+ R^\intercal\right\} \tag{D.1}$$

by showing that the product $MM^{-1} = I_n$, $M^{-1}M = I_n$, (Proposition 80), with the help of $QQ^{+}R^{\intercal} = R^{\intercal}$ (Proposition 76). Observe that directly inverting matrix M is a task of inverting a $n \times n$ matrix while computation of (D.1) requires at most to compute the Moore-Penrose inverse of a $k \times k$ matrix, $k << n$.

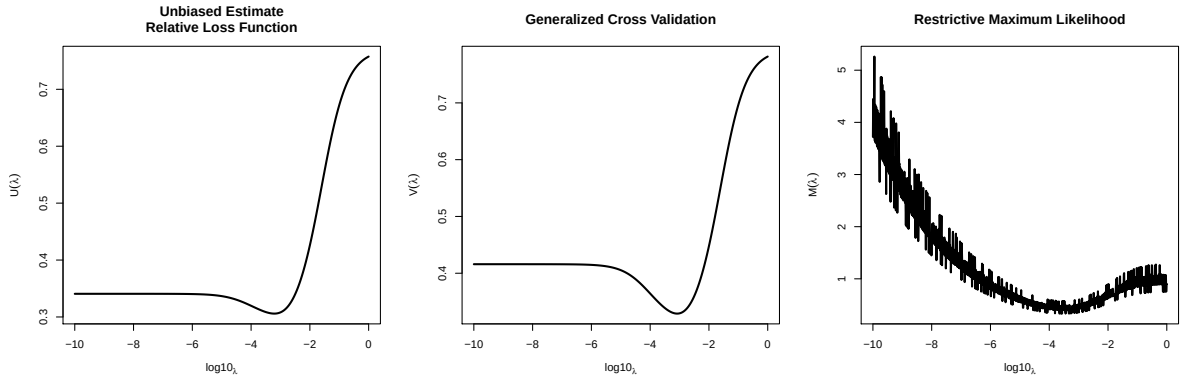

Figure D.1 Plots of the scores used to choose the smoothing parameters computed by their definitions (2.55), (2.59) (2.72). Observe the evident instability for the Restrictive Maximum Likelihood score. The three scores are slow to evaluate if they are computed from their definition.

The inverse of $S^{\intercal}M^{-1}S$ is desired to be computed as well. Expression $\left(S^{\intercal}M^{-1}S\right)^{-1}$ can be more efficiently computed by first obtaining the Cholesky decomposition, (Golub and Van Loan (2012)), of the left side of equation (2.51) as

$$\begin{pmatrix} S^{\intercal}S & S^{\intercal}R \\ R^{\intercal}S & R^{\intercal}R + n\lambda Q \end{pmatrix} = \begin{pmatrix} G_1^{\intercal} & 0 \\ G_2^{\intercal} & G_3^{\intercal} \end{pmatrix} \begin{pmatrix} G_1 & G_2 \\ 0 & G_3 \end{pmatrix}. \tag{D.2}$$

If R is not full column rank we can pivote the columns of G_3 and write

$$G_3 = \begin{pmatrix} H_1 & H_2 \\ 0 & 0 \end{pmatrix},$$

and define

$$\tilde{G}_3 = \begin{pmatrix} H_1 & H_2 \\ 0 & \delta I \end{pmatrix}$$

for small appropriate $\delta > 0$. If R is full column rank, $\tilde{G}_3 = G_3$. Using this notation we can write

$$\left(S^{\intercal}M^{-1}S\right)^{-1} = \frac{1}{n\lambda}\left\{(S^{\intercal}S)^{-1} + (S^{\intercal}S)^{-1}\, S^{\intercal}R\tilde{G}_3^{-1}\tilde{G}_3^{-\intercal}R^{\intercal}S\,(S^{\intercal}S)^{-1}\right\}. \tag{D.3}$$

Equation (D.3) can be proven to hold by showing directly that $\left(S^\intercal M^{-1}S\right)^{-1}\left(S^\intercal M^{-1}S\right) = I_l$ and $\left(S^\intercal M^{-1}S\right)\left(S^\intercal M^{-1}S\right)^{-1} = I_l$ using (D.3) in one side and plugging in (D.1) in $\left(S^\intercal M^{-1}S\right)$ from the other side. The relations $QQ^+R^\intercal = R^\intercal$, $G_3^\intercal G_3 = R^\intercal\left(I - S(S^\intercal S)^{-1}S\right)R + n\lambda Q$ and $G_3^\intercal G_3\tilde{G}_3\tilde{G}_3^{-\intercal}R^\intercal = R^\intercal$ are needed and would need to be proven (Kim and Gu, 2004, Appendix A).

Expressions (D.1) and (D.3) are useful for the evaluation of the full conditional posteriors and for the scores $\mathcal{U}$ (2.55) and $\mathcal{V}$ (2.59). A efficient evaluation of $\mathcal{M}$ (2.72) is more difficult to achieve but finally one can obtain (Kim and Gu (2004)) that

$$\mathcal{M}(\lambda) \propto \frac{\mathbf{y}^\intercal F_2\left(F_2^\intercal M F_2\right)^{-1}F_2^\intercal\mathbf{y}}{|Q + (n\lambda)^{-1}R^\intercal F_2F_2^\intercal R|_+/|Q|_+} \tag{D.4}$$

where F_2 is from the decomposition of S (2.8).

Using the described alternative expressions we have a faster and more stable evaluation of the scores to select the smoothing parameters, Figure D.2.

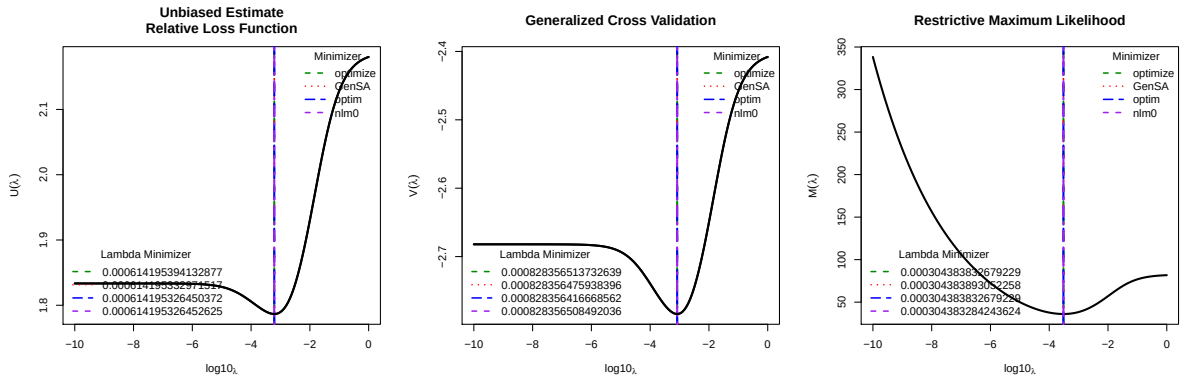

Figure D.2 Plots of the scores using efficient computations of the matrices M^{-1} and $(S^\intercal M^{-1}S)^{-1}$. The score functions are faster to evaluate and numerical instability is no longer present.